

Genome Science performed with e-Science Tools

R.D. Stevens,^a H.J. Tipney,^b C.J. Wroe,^a T.M. Oinn,^c
M. Senger,^c P.W. Lord,^a C.A. Goble,^a A. Brass,^a
M. Tassabehji^b

^aDepartment of Computer Science
University of Manchester
Oxford Road
Manchester
United Kingdom M13 9PL

^bUniversity of Manchester
Academic Unit of Medical Genetics
St Mary's Hospital
Hathersage Road
United Kingdom M13 0JH

^cEuropean Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
United Kingdom CB10 1SD

Abstract

In this paper we describe the use of ^{my}Grid middleware services to create and manage the information from running *in silico* bioinformatics experiments in a semantically enriched Grid aware environment. Through the development and application of workflows using ^{my}Grid components in the context of mapping the complex genomic region associated with the sporadically occurring genetic disorder Williams-Beuren Syndrome, scientists have not only produced biologically interesting and valid results, but also dramatically improved their productivity by undertaking repetitive tasks significantly faster and with an increased diligence compared to previous manual undertakings. In this work we have demonstrated the utility of an e-Science approach to the management of *in silico* experiments.

1 Introduction

This paper describes the use of the ^{my}Grid middleware (Stevens et al., 2003) services to create and manage the information from running *in silico* bioinformatics experiments in a semantically enriched Grid aware environment. This is done in the context of Williams-Beuren Syndrome, a microdeletion in a complex region of human chromosome 7 (Morris, 1988), which requires repeated application of a range of standard bioinformatics techniques to characterise the region deleted in the syndrome and produce a complete genetic map (Stevens et al., 2004).

Bioinformatics already offers a huge selection of data and analytical resources for a biologist to perform *in silico* experiments. In such experiments, services representing tools act upon data, producing more data until a goal is achieved or hypothesis revealed. With current tools it is already possible to reveal interesting biological insights computationally.

A major barrier, however, in utilising these resources is the time needed by skilled bioinformaticians to manually and repeatedly co-ordinate multiple tools to produce a result. Tasks that take minutes of computational time, actually take days to run manually, usually via Web based tools that necessitate much cutting and pasting of data. The aim, therefore, in this work is to improve the manner in which *in silico* bioinformatics experiments are performed and managed, while still using the existing tools.

Performing a series of analyses by hand using Web based tools can be a long, tedious, error-prone (due to slips) and unattractive process. In addition, results, particularly intermediate results, are difficult to record and co-ordinate in a sustainable fashion. Files are often saved to a file store using arbitrary naming schemes in an unsystematic fashion. Lab books record partial details of experiments and it is often easier to re-run the whole experiment than to look at data recorded at the time of the experi-

ment. One aim of e-Science tools should be to enable good practice in the performance of *in silico* experiments.

The increasing importance given to the results of bioinformatics analyses by research groups makes the manual approach increasingly untenable:

1. Many bioinformatics experiments involve a large number of steps. Performing these steps by hand is time consuming, often mundane, and so liable to error.
2. Information is added to public databases at an increasingly fast rate. Bioinformatics experiments should be re-run regularly in order to quickly detect relevant novel sequences.
3. When performed by hand, much of the knowledge on how to perform the bioinformatics experiment remains undocumented and there is a great deal of reliance on expert bioinformaticians.
4. Repetitively performing complex experiments quickly produces large amounts of inter-related data. It becomes difficult to record the origin of large numbers of data files by hand. Good scientific practice demands that such experiments can be replicated, re-run and recorded in such a way to allow good verification and validation.

The field of e-Science promises to utilise current advances in software infrastructure, such as The Grid, to support scientists with their greater reliance on computational methods and to tackle the requirements outlined above. *myGrid* is a UK e-Science pilot project which is developing Grid middleware infrastructure specifically to support *in silico* experiments in biology. From the issues facing the scientist come a strong set of requirements to automate features of the experimental process, its repetition and also support the management of the results.

myGrid addresses these requirements by regarding *in silico* experimental materials and methods as workflows (Stevens et al., 2003). These workflows automate the running of experiments by orchestrating the services that process data. *myGrid* not only supports the creation of the experimental protocol (the workflow), but also the management of the inputs, outputs, intermediates, hypotheses, findings and process enactment records; for the individual and wider groups of scientists. This includes an awareness of the experiments and data holdings of the user, his or her colleagues and the wider scientific community. The aim is to place the scientist at the centre of a virtual bioinformatics organisation and provide the flexibility of data management that affords that scientist a *personalised* view of his or

her data. This information is required if the scientist is to question “How was that result derived?”, “What results have I reviewed so far and which need further investigation?”, “How many times has this experiment been run?”, *etc.* .

We demonstrate the *myGrid* approach in the context of a series of bioinformatics experiments focused on a 1.5 Mb region on Chromosome 7 which is deleted in Williams-Beuren Syndrome (WBS) (see Section 2). Due to the highly repetitive nature of sequence flanking/in the WBS Critical Region (WBSCR), sequencing of the region is incomplete leaving documented gaps in the released sequence. *myGrid* was used in a series of experiments to find newly sequenced human genomic DNA clones that extended into these “gap” regions in order to produce a complete and accurate map of the WBSCR. Once placed in this region, these DNA sequences were analysed with a battery of prediction tools in order to locate putative genes and regulatory elements possibly implicated in the disorder. Finally, any genes discovered were submitted to a range of standard bioinformatics tools for their characterisation. We report how *myGrid* has been used to create workflows for these *in silico* experiments, run those workflows regularly and notify the biologist when new DNA and genes are discovered. The *myGrid* services collect and co-ordinate data inputs and outputs for the experiment, as well as much provenance information about the performance of experiments on Williams-Beuren Syndrome. In this work, *myGrid* has demonstrated the use of output from the UK e-Science programme to answer real questions within biology. The approach has also greatly changed how collaborating biologists perform their *in silico* experiments. The single greatest change is the reduction of a two day manual task to an automated one that can take as little as 30 minutes to run. All data produced by the tools used are systematically recorded and co-ordinated, allowing the scientist to review, verify and validate his or her results and draw the appropriate conclusions. Results management, together with systematic coverage of all the data by all the tools increases the diligence of the performance of the *in silico* experiments. this task is changed from one that runs concurrently with individual component use to a *post hoc* task. the workflows created can be re-run easily and re-purposed to change the analysis performed.

2 Williams-Beuren Syndrome

Williams-Beuren Syndrome (WBS)¹ is a rare, sporadically occurring microdeletion disorder charac-

¹OMIM: #194050

terised by a unique set of physical and behavioural features (Morris, 1988). WBS is caused by a 1.5 Mb deletion (Osborne et al., 2001) located in chromosome band 7q11.23 (Ewart et al., 1993). WBS is a complex, multisystem genetic disorder with an intricate phenotype (Ewart et al., 1993; Osborne, 1999; Preus, 1984). The region commonly deleted in WBS is flanked by highly repetitive regions, \approx 320-500 kb in length (Peoples et al., 2000) containing both pseudogenes and genes.

Most WBS individuals have a deletion of \approx 1.5Mb, encompassing 24 genes (Tassabehji, 2003) (see Figure 1). A smaller region within the common WBS deleted region containing the genes whose absence are critical to the WBS (the Williams-Beuren Syndrome Critical Region) phenotype has been identified (Osborne, 1999; Tassabehji et al., 1999) (see Figure 1).

Many maps of the region have been published (DeSilva et al., 1999; Peoples et al., 2000; Valero et al., 2000; Osborne et al., 2001), each with an increasing level of detail, and the “complete” chromosome 7 sequence was released in 2003 (Hillier, 2003), but still a fully comprehensive map of the WBS region is not available. The overriding reason for this is the complexity and repetitive nature of the WBS region which has led to inconsistencies between published maps and hard to close gaps in the genomic sequence.

The gaps in the WBS region may harbour important genes and associated regulatory elements which are deleted; so defining their composition is crucial for genotype-to-phenotype correlations. The production of a complete, comprehensive and robust map of the WBS region is vital if we are to fully understand the pathology of WBS. It is the creation of such a map that ^{my}Grid is used to create and run a series of *in silico* experiments.

3 Williams-Beuren Syndrome Bioinformatics

To tackle the creation of a genetic map, two separate bioinformatics analyses are performed, usually by hand:

1. Retrieve newly submitted human genomic sequences that extend into the gap. To do this, similarity searches (used to find matching regions of DNA) are made against a range of GenBank databanks using the BLAST programme BLASTN (Altschul et al., 1997). Repeat Masker²

²<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>

is used to search against RepBase Update 6.3 (Jurka, 2000) to avoid spurious multiple hits against repetitive sequences (see Figure 2 for details).

2. Any high-scoring matches from human chromosome 7 are submitted to the NIX programme³ which is used to find any gene(s) residing on those new fragments. To characterise any gene(s), surrounding regions and any putative gene product(s), the genomic DNA is analysed for a full range of motifs and features and translated in all six frames to produce putative protein sequences. The most suitable reading frame is used in a similarity search against protein databanks and submitted to a standard collection of protein characterisation tools.

The ^{my}Grid middleware is a collection of services upon which bioinformatics applications can be built. Each of the services offers solutions for the requirements of the research scientist, either by automating tasks or supporting the overall management of the experiments and their results.

Service provision: To allow automated interaction with bioinformatics applications during the *in silico* experiment we must provide programmatic access to those applications from the workflows that represent the experiment. We achieve this by making each bioinformatics application available as a Web Service⁴. Web Services provide a standardised way of integrating Web-based applications using XML based messaging over an Internet protocol backbone. Many of the applications such as Genscan⁵ and RepeatMasker are available as command line applications. We used Soaplab (Senger et al., 2003), as a framework in order to expose these command line applications as Web Services.

Writing the workflows: Automating the experimental process requires an explicit representation of that process sufficient for a computer to execute. Workflows represent a procedure, such as a bioinformatics analysis, as a set of processes and the relationships between those processes. It is the level of abstraction that is an important aspect of workflows – the user has declarative, rather than procedural access to the analysis. Thus the user describes what he or she wishes to accomplish, not how to accomplish the goal. In this way we aim to allow biologists without programming skills to develop their own *in silico* experiments. The ^{my}Grid team have developed the Simple Conceptual Unified Flow Language (Scufl) and an application to edit workflows (Taverna) (Oinn

³<http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/>

⁴<http://www.w3.org/TR/ws-arch/>

⁵<http://genes.mit.edu/GENSCAN.html>

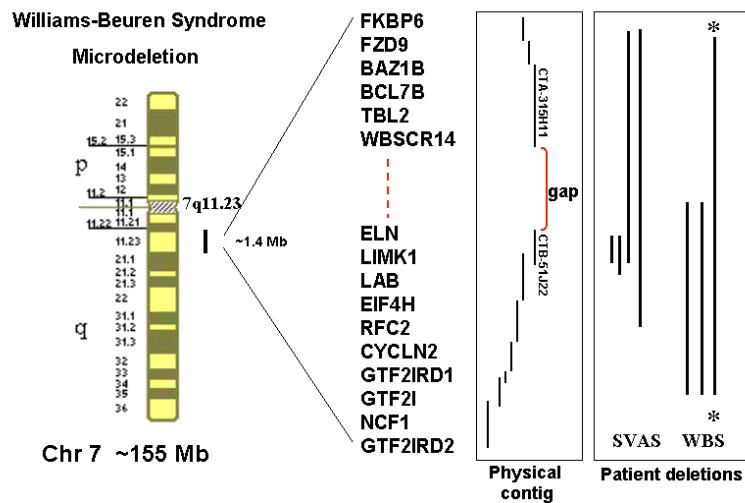


Figure 1: Williams-Beuren Syndrome microdeletions reside on chromosome 7q11.23. Patients with deletions fall into two categories. Those with classic WBS (* indicates the common deletion) and those with SVAS but not WBS, caused by hemizygous deletion of the elastin gene. A physical map of the region composed of genomic clones is shown with a gap in the critical region. The ^{my}Grid software was used to continue the contig and identify more genes at this locus.

et al., 2003) to achieve this abstraction. Thus a biologist or bioinformatician does not need to write a large, bespoke application, but to simply describe what needs to be done; in this way, the analysis workflows were written by the bioinformatician herself. Figure 2 shows a graphical representation of the workflow written to explicitly represent the first manual analysis described above.

Running the workflows: A workflow enactment engine, Freefluo, has been developed for the enactment of workflows written in Scuff (Addis et al., 2003). The engine automatically calls each Web Service in the appropriate order and passes data between services.

Collating the results: Both final and intermediate results from running the workflow are saved. A major requirement is to not only automate the experimental process but also assist the scientist in recording the origin or *provenance* of the large set of inter-related result files. ^{my}Grid addresses these management requirements with several components:

1. A Common Experimental Information Model

At the core is the ^{my}Grid information model which provides a standard by which to structure information about bioinformatics experiments and data. In brief, the model splits into two parts: 1) organisational information such as the members of the research group, data access rights, current projects and their experiments; 2) Information about the life cycle of a single experiment such as its design, when it has been performed, results it has

produced and origin of those results.

We enable the user to explore the context of experimental data by providing associated metadata with each item. This metadata uses a schema derived from the information model and references other items collected life cycle of an *in silico* experiment. Cross referencing between these items requires a common identification system. ^{my}Grid is adopting the Life Science Identifier (LSID) and associated resolution system (Clark et al., 2004). Inputs, intermediate results, final outputs, data about the bioinformatics services, etc. of a workflow are all assigned an LSID by the workflow environment, and stored in a repository which supports the LSID resolution protocol. The protocol also allows for the retrieval of metadata associated with each item, and current implementations for LSID provide this metadata in RDF format (Resource Description Framework). RDF has been developed by W3C to represent structured information on the Web (Klyne and Carroll, 2003). RDF metadata is a collection of statements relating two resources together via an RDF property. In this way it provides a simple graph based model and can represent information in a minimally constraining flexible way. It identifies resources with URI's and so is compatible with LSID.

Figure 3 shows some of the relationships a single item of data can possess, and how these can indirectly link to many entities within the information model, forming a graph or web through which the

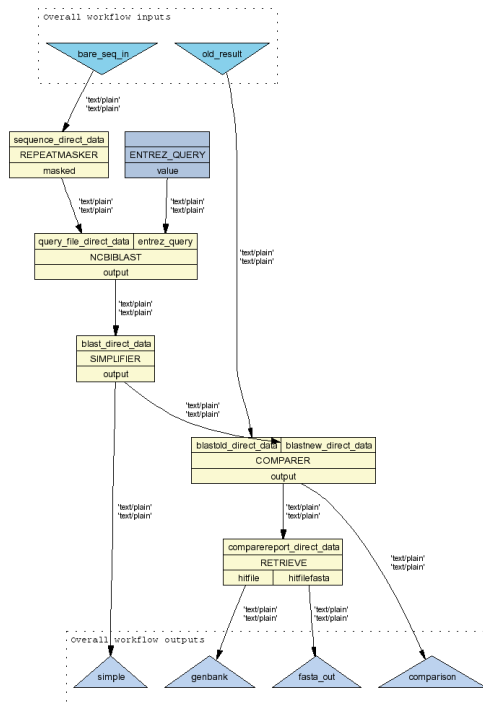


Figure 2: Schematic representation of the first workflow created to explore gap regions within the WBSER. This workflow takes the last verified piece of sequence (< 3000 bp) in the contig flanking a gapped region and produces a shortlist of sequences which may extend the contig into the gap region. The query sequence is masked using RepeatMasker to prevent spurious hits prior to being used by the NCBI BLASTN program to identify overlapping sequences. Only new or improved hits are relevant, and so the results are first translated into a simplified format using “SIMPLIFIER”, before being compared to the results of the previous run using the “COMPARER” service. “RETRIEVE” takes new hits and determines which are located on human chromosome 7 before returning those sequences in FASTA format.

user can navigate to valid results, find other related results and generate different views over his or her body of scientific work.

2. **Automated Provenance Recording** The workflow environment has been built to automatically generate two kinds of metadata. The first is called *process provenance* and is analogous to a log, recording which services were used to generate the data. The second, *data provenance*, provides relationships *between data*. These simply show the derivation path of the data. In most cases these derivation relationships are dependent on specific services, so for example a BLAST service will provide a report which has “similar sequences to” the input query sequence. Therefore each step of the workflow can be annotated with a provenance template which describes the relationship between the data flowing in and out of the process. We describe the knowledge acquired from these templates as *knowledge level* provenance.

The recording of the provenance of actions taken in “joining” services, together with the storage of

intermediate results, means that each run of an experiment can be fully validated by a user “tracing” back through the co-ordinated set of results.

Notification: As the experiment run is automatic, it is important for the system to be proactive and notify the scientists when new results have been produced. In this case when new sequences on chromosome 7 have been matched. myGrid has a notification service to provide such asynchronous messaging (Krishna and et al, 2003). Services may register the type of notification events they produce and clients may register their interest in receiving updates. The type and granularity of notification events is defined with ontological descriptions in metadata supplied to the notification service. For example, the scientist can subscribe to notifications of changes in the results of this workflow.

Viewing the results: As much of the information has been recorded by machine it must be rendered in a human readable form. The amount and complexity of the information also means that it must be provided in filtered views that help answer specific

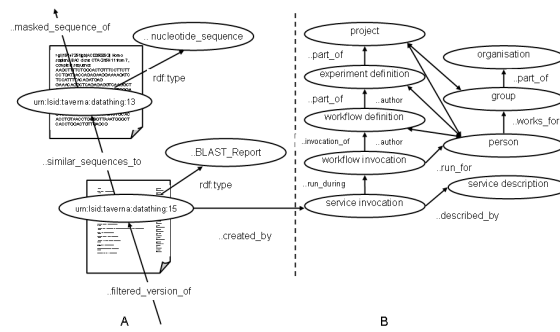


Figure 3: (A) shows the relationships data item (in this case a BLAST report) can have with other data items in the repository. Each item is represented as an RDF resource identified using an LSID. RDF statements (a triple of two nodes and an arc) are automatically generated to relate (in this case) the BLAST report with the query sequence used to generate this data, and subsequent filtered versions of the report. Each data item can be typed using concepts from an ontology. In this case a statement is added to explicitly state this is of type `BLAST_Report`. (B) shows how the data item can be related to other classes of information in the information model schema. The BLAST report has been generated by a service invocation which in turn is part of a larger workflow invocation. The workflow has been run for a particular person based on a workflow definition. The workflow forms part of an *in silico* experiment used by a research group as part of a specific project.

questions clearly (such as those stated in the introduction). To achieve this we have used Haystack, a desktop application that allows users to browse multiple views of RDF based information. Its previous use within the bioinformatics area (Quan and Karger, 2004) has allowed us to reduce the time needed to tailor the application for this project. Using Haystack, the user is able to navigate a graph of relationships similar to that shown in Figure 3. If she needs to focus in on a particular item in more detail she can drill down and display a summary page.

4 Results: Newly Extended Williams-Beuren Syndrome Region

The success of the workflows created for the WBS experiments in *myGrid* is apparent on a number of levels. Biologically interesting and correct results were achieved, with results gathered from more than one iteration of the workflows for the WBS experiments enabling significant extension of the centromeric WBS CR contig. In the first instance, BAC RP11-622P13 (gi:15145617, gb: AC073846) was identified as overlapping the centromeric end of the WBS CR contig and extended this contig by 121004 bp. Within this 'new' sequence six putative coding sequences (genes) were identified; five of which were identified as corresponding to the five known genes in this region. CDS1 correctly identified all six known exons of WBS CR21 variant 1 (gi: 23200007, gb: NM_148912), CDS3 correctly identified nine exons (including the PolyA-tail) out of the

ten known to reside within STX1A (gi:4759181, gb: NM_004603), CDS4 correctly identified all 12 exons and the PolyA-tail of WBS CR22 (gi: 23199994, gb: NM_017528), CDS5 correctly identified the single known exon of WBS CR18 (gi: 22538496, gb: NM_032317), and CDS6 partially identified WBS CR24 (gi: 37539029, gb: XM_353620) correctly identifying 3 of its 5 exons. Having extended the contig it was possible to search for the next overlapping region of sequence by applying the last 3000 bp from the telomeric end of BAC RP11-622P13 to our workflow. From this second application of the process, two putative genomic clones were identified and further examination identified BAC RP11-148M21 (gi:18450186, gb: AC093168) as closing the gap by a further 146 689 bp. Five putative coding sequences were predicted; two of which correctly identified in full the two known genes in this region. CDS1 correctly identified the single exon of CLDN3 (gi:21536298, gb: NM_001306), while CDS2 precisely identified the single exon gene CLDN4 (gi:34335232, gb: NM_001305). In summary, just two iterations of our workflows taking approximately one day to run, correctly reduced this gap by 267 693 bp at its centromeric end, correctly located all seven known genes in this region and identified 33 of the 36 known exons residing in this location. These results have been confirmed by repeating experiments using conventional manual interaction with the relevant Web sites and previous knowledge of sequences that extend into this particular gap. We fully expect the refined and optimised *myGrid* workflows to reveal new results as they are run on a regular basis.

5 Discussion

^{my}Grid has been successful in capturing the *in silico* experimental requirements of its biologists. Workflows representing the bioinformatics analyses have been written, by the biologist herself, run within the workflow environment and the provenance data captured, co-ordinated and browsed by the biologist to validate her results. Genomic DNA that had already been analysed was used to verify that the workflow found and collected the appropriate bioinformatics findings and these can be seen in Section 4.

A finding of equivalent importance to the results themselves is the impact these workflows have had on the way the authors work. Manually, the processes undertaken by the workflows developed here could take at least two days on Web based tools, while the workflows achieve the same output in approximately 30 minutes. This has a significant impact on the productivity of the scientist, especially when considering these experiments are often undertaken weekly, enabling the experimenter to act on interesting information quickly without being bogged down with the monitoring of services and their many outputs as they are running. In a more social aspect, complex Web based working is intricate and tedious. It is a long task that can be endlessly delayed due to its unattractiveness. Automating the process removes this barrier.

For a given piece of genomic DNA as the original input the workflow, each of the results from an individual Web Service are all sent as input to the next Web service bioinformatics tool. This systematic approach is cheap computationally and increases the diligence of the *in silico* experiment. The system also enables the scientist to view all the results at once, selecting those which appear to be most promising and then looking back through the results to identify areas of support, again increasing the diligence of the process. This is quite different to the ongoing analysis of all results, correct or otherwise, produced by each service – the case when undertaking such a task manually. Thus, the workflow approach has moved results analysis from being concurrent with results production to being a completely *post hoc* activity.

The core of these findings, that an automatic analysis has efficiency and effectiveness advantages over a manual task are not new. Bioinformaticians have been capturing complex analyses within bespoke programmes, often PERL scripts, as long as the discipline has existed. So what are the advantages of using a workflow based approach? Our biology collaborators are not bioinformatics application builders. These WBS workflows could have been implemented by or for the biologists as a bespoke

programme. If created by the biologists themselves, who are not programmers, the task would have been greatly extended. If a bespoke application had been built by commission, it is less likely that the biology requirements would have been met so easily. The high-level, declarative nature of the workflow construction meant that the biologist herself was easily able to create the workflows without too much intervention. This enabled her to build exactly what she wanted to have in the analyses.

Again, a bespoke application could have the results management built into that application, but this all comes for free in the ^{my}Grid approach. The materials and method of the analysis are automatically recorded and co-ordinated in a consistent fashion. The separation of workflow environment and results presentation means that third party solutions, as long as they are RDF compliant, can be used over the data sets produced by ^{my}Grid.

Finally, workflows aim to promote re-use and repurposing. The nature of their creation means that adding or replacing services can be accomplished with relative ease. Since the workflow shown in Figure 2 was first created, it has been extended and enhanced to provide more bioinformatics results. These help characterise the putative proteins produced by the genes and further characterise the DNA encoding those genes. These new results from the enhanced workflows are described and co-ordinated in the same way as the previous results and through this annotation are linked together forming a web of scientific data holdings on Williams-Beuren Syndrome.

This collaboration has shown ^{my}Grid, a UK e-Science pilot project, significantly contributing to a scientific investigation and changing work-practices for the better. ^{my}Grid is planning further collaborations to perform more real science and we expect to have a concomitant feedback into the development of ^{my}Grid.

Acknowledgements: The authors would like to acknowledge the assistance of the whole ^{my}Grid consortium. This work is supported by the UK e-Science programme EPSRC grant GR/R67743. MT and HT are supported by The Wellcome Foundation (G/R:1061183).

References

- Addis, M., Ferris, J., Greenwood, M., Li, P., Marvin, D., Oinn, T., and Wipat, A. (2003). Experiences with eScience workflow specification and enactment in bioinformatics. In *Proc UK e-Science All Hands Meeting 2003*, pages 459–466.
- Altschul, S. F., Thomas, L., Madden, A., Schaffer,

- A., Zhang, J., Zhang, Z., Miller, W., and Lipman, David, J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402.
- Clark, T., Martin, S., and Liefeld, T. (2004). Globally Distributed Object Identification for Biological Knowledgebases. *Briefings in Bioinformatics*, 5(1):59–70.
- DeSilva, U., Massa, H., Trask, B., and Green, E. (1999). Comparative mapping of the region of human chromosome 7 deleted in Williams Syndrome. *Genome Research*, 9:428–436.
- Ewart, A., Morris, C., Atkinson, D., Jin, W., Sternes, K., Spallone, P., Stock, A., Leppert, M., and Keating, M. (1993). Hemizygoty at the elastin locus in a development disorder, williams syndrome. *Nature Genetics*, 5:11–16.
- Hillier, L. (2003). *et al.* The DNA Sequence of human chromosome 7. *Nature*, 242:157–164.
- Jurka, J. (2000). Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics*, 9:418–420.
- Klyne, G. and Carroll, J. J. (2003). Resource description framework (RDF): Concepts and abstract syntax. *W3C Proposed Recommendation*. Available at <http://www.w3.org/TR/2003/PR-rdf-concepts-20031215/>.
- Krishna, A. and et al (2003). myGrid Notification Service. In *Proc UK e-Science All Hands Meeting 2003*.
- Morris, C. (1988). The natural history of Williams syndrome: physical characteristics. *Journal of Paediatrics*, 113:318–326.
- Oinn, T., Addis, M., Ferris, J., Marvin, D., Greenwood, M., Carver, T., Pocock, M. R., Wipat, A., and Li, P. (2003). Taverna: A Tool for the Composition and Enactment of Bioinformatics Workflows. Accepted for publication.
- Osborne, L. (1999). Williams-Beuren Syndrome: Unraveling the mysteries of a microdeletion disorder. *Molecular Genetics and Metabolism*, 67:1–10.
- Osborne, L., Li, M., Pober, B., Chitayat, D., Boddurtha, J., Mandell, A., Costa, T., Grebe, T., Cox, S., Tsui, L.-C., and Scherer, S. (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics*, 29:321–325.
- Peoples, R., Franke, Y., Wang, Y.-K., Pérez Jurado, L., Paperna, T., Cisco, M., and Francke, U. (2000). A Physical map, including BAC/PAC clone contig, of the Williams-Beuren Syndrome-deletion region at 7q11.23. *American Journal of Human Genetics*, 66:47–68.
- Preus, M. (1984). The Williams syndrome: objective definition and diagnosis. *Clinical Genetics*, 25:422–428.
- Quan, D. and Karger, D. R. (2004). How to make a semantic web browser. In *Thirteenth International World Wide Web Conference*, pages 284–285, New York, USA.
- Senger, M., Rice, P., and Oinn, T. (2003). SoapLab a unified Sesame door to analysis tools. In *Proc UK e-Science All Hands Meeting 2003*.
- Stevens, R., Tipney, H., Wroe, C., Oinn, T., Senger, M., Lord, P., Goble, C., Brass, A., and Tassabehji, M. (2004). Exploring Williams Beuren Syndrome Using myGrid. Accepted for publication at Intelligent Systems for Molecular Biology (ISMB) 2004.
- Stevens, R. D., Robinson, A. J., and Goble, C. A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19:i302–i304.
- Tassabehji, M. (2003). Williams-Beuren syndrome: a challenge for genotype-phenotype correlations. *Human Molecular Genetics*, 12:R229–R237.
- Tassabehji, M., Metcalfe, K., Karmiloff-Smith, A., Carette, M., Grant, J., Dennis, N., Reardon, W., Splitt, M., Read, A., and Donnai, D. (1999). Williams syndrome: Use of chromosomal microdeletions as a tool to dissect cognitive and physical phenotypes. *American Journal of Human Genetics*, 64:118–125.
- Valero, M., de Luis, O., Cruces, J., and Prez Jurado, L. (2000). Fine-scale comparative mapping of the human 7q11.23 region and the orthologous region on mouse chromosome 5G: The low-copy repeats that flank the Williams-Beuren syndrome deletion arose at breakpoint sites of an evolutionary inversion(s). *Genomics*, 69:1–13.