BIOINFORMATICS

Vol. 20 Suppl. 1 2003, pages i303–i310
DOI: 10.1093/bioinformatics/bth944

Exploring Williams–Beuren syndrome using myGrid

R. D. Stevens[1,*], H. J. Tipney[2], C. J. Wroe[1], T. M. Oinn[3], M. Senger[3], P. W. Lord[1], C. A. Goble[1], A. Brass[1] and M. Tassabehji[2]

[1]Department of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK, [2]University of Manchester, Academic Unit of Medical Genetics, St. Mary's Hospital, Hathersage Road, M13 0JH, UK and [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: *In silico* experiments necessitate the virtual organization of people, data, tools and machines. The scientific process also necessitates an awareness of the experience base, both of personal data as well as the wider context of work. The management of all these data and the co-ordination of resources to manage such virtual organizations and the data surrounding them needs significant computational infrastructure support.

Results: In this paper, we show that myGrid, middleware for the Semantic Grid, enables biologists to perform and manage *in silico* experiments, then explore and exploit the results of their experiments. We demonstrate myGrid 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). 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 characterization. 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 WBS.

Availability: The myGrid software is available via http://www.mygrid.org.uk

Contact: robert.stevens@cs.man.ac.uk

*To whom correspondence should be addressed.

1 INTRODUCTION

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 possible to reveal interesting biological insights computationally. A major barrier, however, in utilizing 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. This paper describes the use of the myGrid 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 (WBS), a microdeletion in a complex region of human chromosome 7, which requires repeated application of a range of standard bioinformatics techniques to characterize the region deleted in the syndrome.

Due to the highly repetitive nature of the sequence flanking/in the Williams–Beuren syndrome critical region (WBSCR), sequencing of the region is incomplete leaving documented gaps in the released genomic sequence. In order to produce a complete and accurate map of the WBSCR, researchers must constantly search for newly sequenced human DNA clones which extended into these 'gap' regions (see Section 3). Once placed in this region, these DNA sequences must be analysed with a battery of prediction tools to locate putative genes, their regulatory elements, as well as both characterized and otherwise uncharacterized genes and their products implicated in WBS.

As part of ongoing efforts to produce a complete map of the WBSCR, the authors would have historically closed these gaps by hand, manually and repeatedly interacting with a range of standard bioinformatics services on the Web until enough information had been gathered to characterize the gap region. Results from one task would have been manually copied to form the input of another task. Each time an individual embarks upon this process a large set of results files are saved to their local file system, in addition to the origin, relevance and current status of each file being recorded in their hand written lab book. 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?" and "How many times has this experiment been run?" The greater importance given to bioinformatics results by research groups makes this 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 rerun regularly in order to quickly detect relevant novel sequences. (3) When performed by hand much of the knowledge of 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.

The field of e-Science promises to utilize current advances in software infrastructure, such as The Grid, to support scientists with their greater reliance on computation methods. A Grid is a virtual organization of a set of heterogeneous distributed resources and the networks used to link them. A virtual organization of machines, resources and people is one particular manifestation of a Grid and one that matches the requirements of typical bioinformatics experiments (Foster and Kesselman, 2003).

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, as described above, come a strong set of requirements to automate the experimental process, its repetition and also support the management of the results. myGrid addresses these requirements by regarding *in silico* experiments as workflows (Stevens *et al*., 2003). These workflows automate 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 and findings; 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 organization and the flexibility of data management that affords that scientist a personalised view of his or her data. In this paper we describe the *in silico* experiments required for exploring WBS, and the use of myGrid services to implement and manage the running of those experiments and their results.

2 WILLIAMS–BEUREN SYNDROME

Williams–Beuren syndrome[1] is a rare, sporadically occurring microdeletion disorder characterized 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 $\approx$ 1.5 Mb, encompassing 24 genes (Tassabehji, 2003) (Fig. 1). A smaller region within the common WBS deleted region containing the genes whose absence are critical to the WBS (the WBSCR) phenotype has been identified (Osborne, 1999; Tassabehji *et al*., 1999) (Fig. 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 WBSCR is not available. The overriding reason for this is the complexity and repetitive nature of the WBSCR which has lead to inconsistencies between published maps and hard to close gaps in the genomic sequence.

The gaps in the WBSCR 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.

3 WILLIAMS–BEUREN SYNDROME BIOINFORMATICS IN MYGRID

The biological problem described above has previously been investigated manually in two major analyses:

1. Retrieve newly submitted human genomic sequences that extend into the gap. Similarity searches are made against a range of GenBank databanks using the BLAST programme BLASTN (Altschul *et al*., 1997). Repeat Masker[2] is used to search against RepBase Update

[1] OMIM: #194050.

[2] http://www.repeatmasker.org.

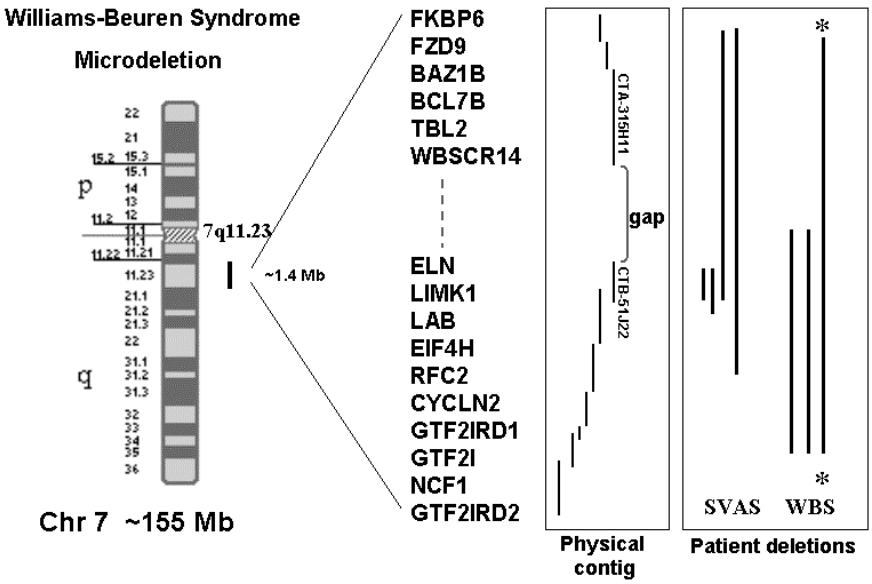

Fig. 1. Williams–Beuren syndrome microdeletions reside on chromosome 7q11.23. Patients with deletions fall into two categories. Those with classic WBS (* indicates the common deletion) and those with SVAS but not WBS, caused by hemizygous deletion of the elastin gene. A physical map of the region composed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

6.3 (Jurka, 2000) to avoid spurious multiple hits against repetitive sequences (Fig. 2 for details).

2. Any high-scoring matches from human chromosome 7 are submitted to the NIX programme[3] which is used to find any gene(s) residing on those new fragments. To characterize 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. The most suitable reading frame is used in a similarity search against protein databanks and submitted to a standard collection of characterization tools.

In order to transfer this manual procedure into the myGrid environment we had to go through a series of steps outlined below. Each step addresses the requirements of the research scientist, either by automating tasks or supporting the overall management of the experiments and their results.

(1) *Service provision.* To allow automated interaction with applications during the *in silico* experiment we must provide programmatic access to those applications. We achieve this by making each bioinformatics application available as a Web Service[4]. Web Services provide a standardized way of integrating Web-based applications using XML-based messaging over an Internet protocol backbone. Many of the applications such as Genscan[5] 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. Grid Services promise additional functionality and myGrid has planned a migration path to the Open Grid Services Architecture (Foster *et al*., 2002).

(2) *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. The myGrid team have developed the Simple Conceptual Unified Flow Language (Scufl) and an application to edit workflows (Taverna Scufl Workbench) (Addis *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 graphical representations of the workflow written to explicitly represent the first manual analysis described above. The workflow uses only two main bioinformatics applications RepeatMasker and NCBI BLAST. As can be seen from the workflow, however, many other 'joining' services are required to produce the desired results. For example, the service labelled "RETRIEVE" takes new hits from a BLAST search; retrieves the GenBank record; determines which lie upon chromosome 7; and then reformats the sequence as a

[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.

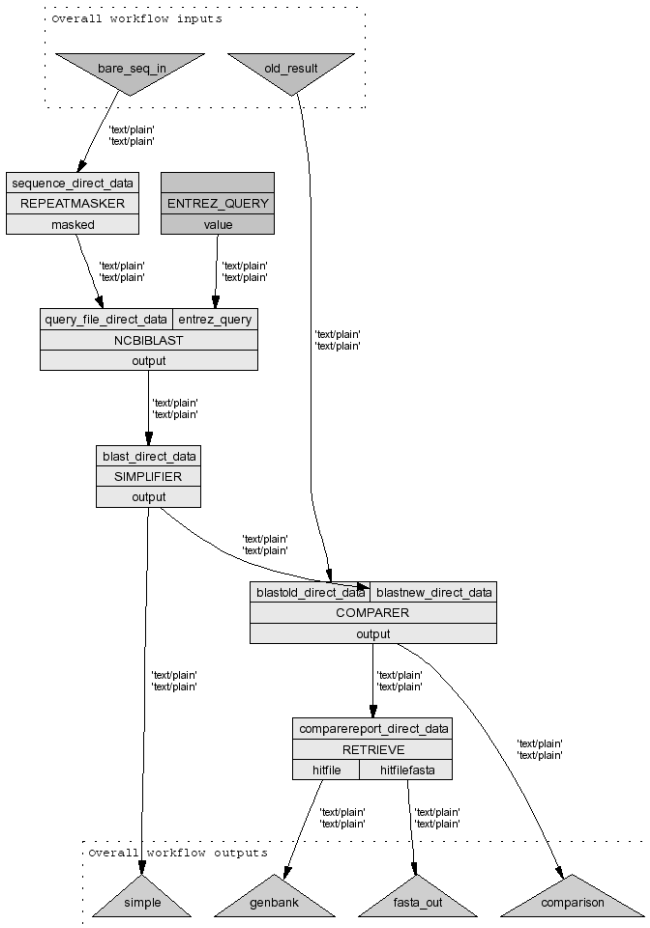

Fig. 2. Schematic representation of the first workflow created to explore gap regions within the WBSCR. This workflow takes the last verified piece of sequence (<3000 bp) in the contig flanking a gapped region and produces a short list of sequences which may extend the contig into the gap region. The 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. Intermediate results are kept in case the filtering operation has excluded relevant but mis-annotated sequences.

FASTA file. Often many of these hidden steps performed by the scientist remain undocumented, but become explicit in the workflow. Several iterations are required in writing the workflow to ensure it accurately reflects what the scientist achieves by hand. This not only makes automation possible but also promotes sharing of experimental knowledge. Tacit expertise held by an individual is thus now accessible as a formally documented procedure.

(3) *Running the workflows.* A workflow enactment engine, Freefluo, has been developed for the enactment of workflows written in Scufl (Addis *et al*., 2003). The engine automatically calls each service in the appropriate order and passes data between services. For simplicity the Taverna Scufl workbench can be used to send the workflow script and required data to the enactment engine. However, a more comprehensive myGrid bioinformatics workbench is also under development which includes interfaces to the myGrid information repository (mIR) capable of storing and managing bioinformatics data for a distributed group of researchers (Goble *et al*., 2003).

(4) *Collating the results.* Both final and intermediate results from running the workflow are saved either in the users local file system or mIR. 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. myGrid addresses these management requirements with several components:

(a) *A common experimental information model.* At the core is the myGrid 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) organizational 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 provenance of those results.

The information model is intended to pervade the myGrid architecture at all levels. The mIR stores information corresponding to entities in the model. XML messages between myGrid Web Services are being designed to conform to the information model. Finally, 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 in the repository. Crossreferencing between these items requires a common identification system. myGrid is adopting the life science identifier (LSID) and associated resolution system (Clark and Liefeld, 2004). LSIDs are a special kind of universal resource name (URN) as developed by W3C consortium. They have their own resolution protocol, which provides for persistent location independent object identifiers. Inputs, intermediate results and final outputs of a workflow are all assigned an LSID by the workflow environment, and stored in the 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 resource description framework format (RDF). RDF has been developed by members of W3C to represent structured information on the Web (Klyne

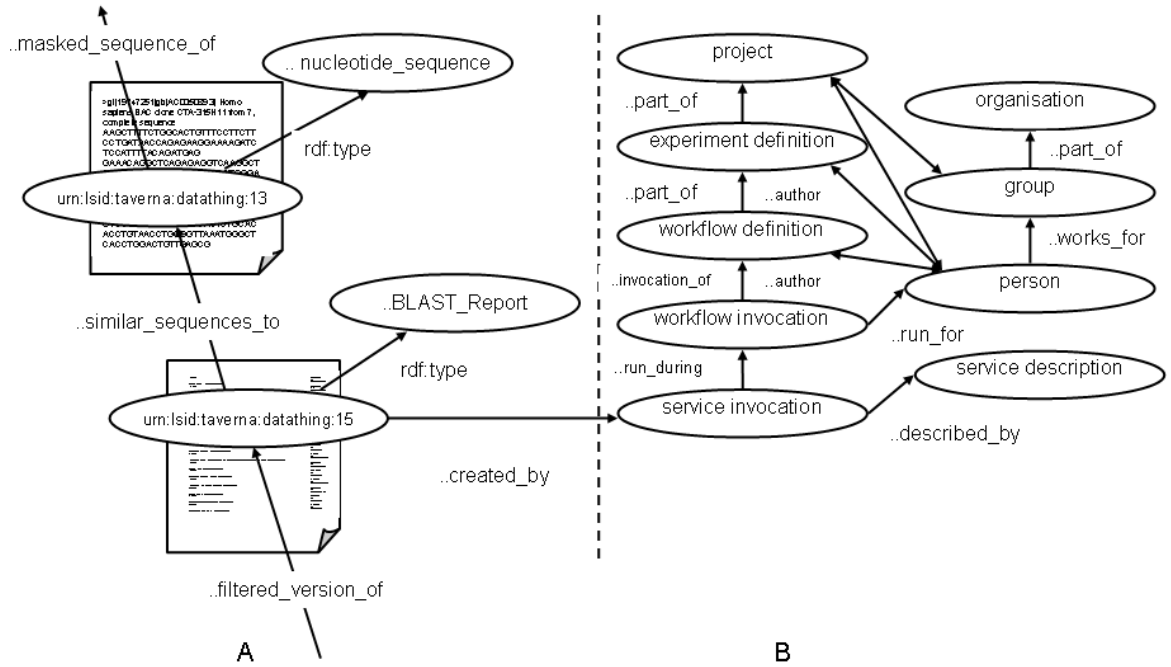

Fig. 3. (**A**) shows the relationship data item (in this case a BLAST report) can have with other data items in the repository. Each item is represented as an RDF resource identified using an LSID. RDF statements (a triple of two nodes and an arc) are automatically generated to relate (in this case) 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.

and Carroll, 2003, http://www.w3.org/TR/2003/PR-rdf-concepts-20031215/). 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 URIs (of which URNs are a type), and so is compatible with LSID. It provides explicit links to ontologies written in standard languages such as the Web Ontology Language (OWL)[6] allowing resources to be typed using ontological concepts.

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

(b) *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 provides relationships between data. In most cases these are dependent on specific services, so e.g. 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. 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.

(5) *Notification.* As the experiment can run unattended, 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 *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,

[6] http://www.w3.org/TR/2004/Rec-owl-features-20040210/.

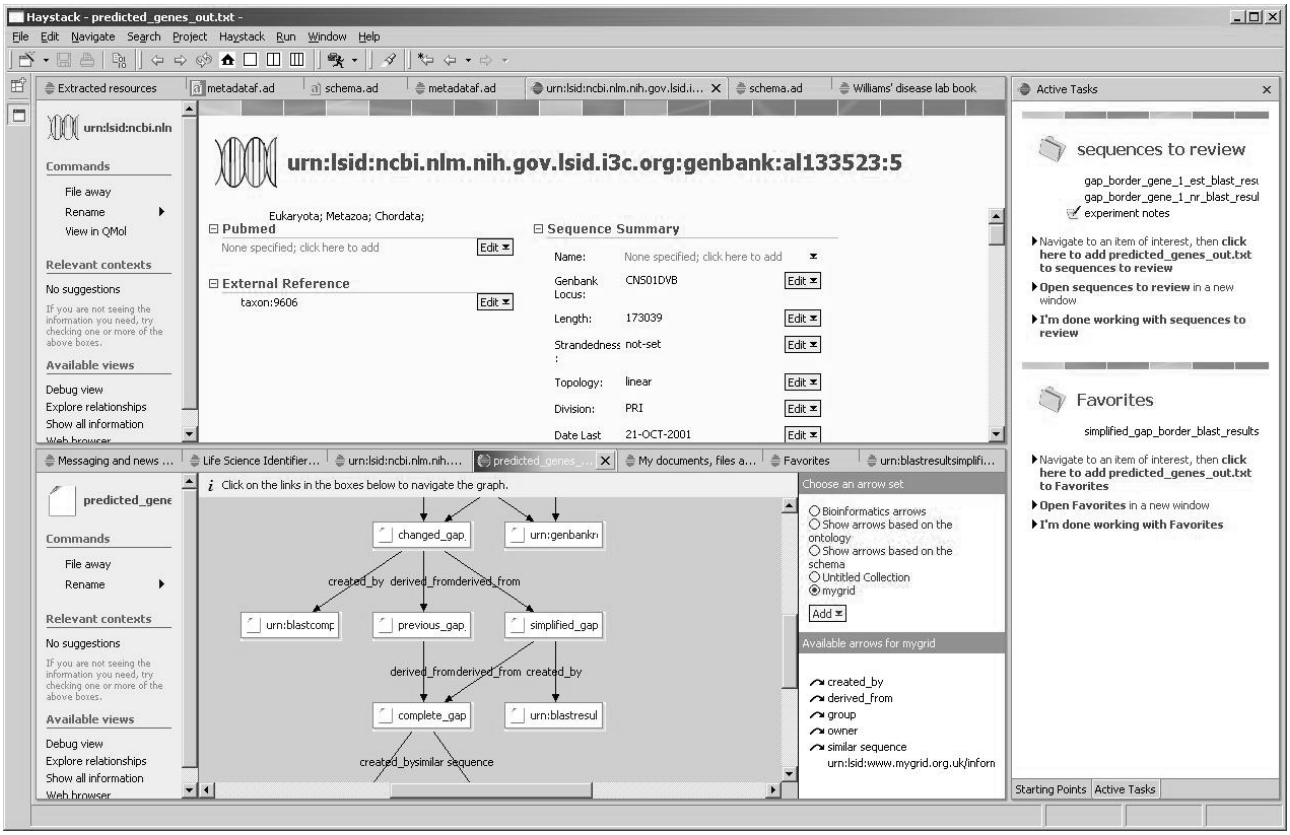

Fig. 4. Screen shot of Haystack showing several views of information produced from the first experimental run. The top central pane shows a view of a GenBank record displaying information about a relevant sequence. The bottom central pane shows a portion of the web of relationships indicating where this result originated. The right hand pane shows the current active user task of managing the collection of sequences for review. Relevant results can be dragged on and off this list as appropriate.

the scientist can subscribe to notifications of changes in the results of this workflow.

(6) *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 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 *et al.*, 2003, http://www.ai.mit.edu/people/dquan/iswc2003-bioinformatics.pdf) has allowed us to reduce the time needed to tailor the application for this project. Figure 4 shows the use of Haystack to examine information regarding the initial run of the experiment. 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 for that resource. In the example shown in Figure 4, Haystack is showing a summary of a Genbank record.

4 RESULTS: NEWLY EXTENDED WILLIAMS–BEUREN SYNDROME REGION

The success of the workflows created for the WBS experiments in ^{my}Grid 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 WBSCR contig. In the first instance, BAC RP11-622P13 (gi:15145617, gb: AC073846) was identified as over lapping the centromeric end of the WBSCR contig and extended this contig by 121 004 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 WBSCR21 variant 1 (gi: 23200007, gb: NM_148912), CDS3 correctly identified nine exons (including the PolyA-tail) out of the 10 known to reside within STX1A (gi:4759181, gb: NM_004603), CDS4 correctly identified all 12 exons and the PolyA-tail of WBSCR22 (gi: 23199994, gb: NM_017528), CDS5 correctly identified the single known exon of WBSCR18 (gi: 22538496,

gb:NM_032317) and CDS6 partially identified WBSCR24 (gi: 37539029, gb: XM_353620) correctly identifying three of its five 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 precisley identified the single exon gene CLDN4 (gi:34335232, gb: NM_001305). In summary, just two iterations of our workflows taking approximately 1 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 optimized myGrid workflows to reveal new results as they are run on a regular basis.

A finding of equivalent importance 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 2 days, while the workflows achieve the same output in approximately an hour. This has a significant impact on the productivity of the scientist, especially when considering these experiments are often undertaken weekly, enabling the experimentor to act on interesting information quickly without being bogged down with the monitoring of services and their many outputs as they are running. 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. 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.

5 DISCUSSION

Performing bioinformatics *in silico* experiments entails the orchestration of a large number of distributed resources (machines, tools, databanks and people) into a virtual organization. Such a typical bioinformatics scenario is seen in this investigation of WBS and is the kind of scenario for which Grid technology was developed. No novel bioinformatics procedures have been developed, but the way in which existing procedures have been managed has greatly increased the productivity of the scientists. The workflows developed for this project have applications in other projects. myGrid can be considered a toolbox of components with which a bioinformatician developer can construct targeted applications through a generic mechanism. The next step is therefore to develop a Web application that allows the configuration of these specific workflows and monitoring of their results without any need for the user to directly interact with myGrid.

The necessity for experimental repetition and the complexity of the battery of techniques used in the analysis of new findings involves the generation of a large amount of fragmentary and complex inter-related data holdings. To create awareness of the scientific work context, a scientist needs to be able to explore and manage this experience base. A scientist needs multiple views over these datasets to achieve the following objectives: (i) to determine the areas of work already carried out by the scientist, (ii) to identify work carried out by other groups, (iii) to explore work on a particular topic (e.g. WBS), (iv) to explore from the viewpoint of data objects, such as genes or proteins. myGrid automatically records experimental data holdings as a Web of Science (Hendler, 2003) and allows such personalization over the experience base. Providing a practical application to navigate this Web of Science is however non-trivial. We will continue to work with Haystack in order to refine the presentation of experimental results.

From the bioinformatics viewpoint, experience in formalizing the manual tasks has revealed a significant amount of intermediate steps often not initially made explicit by the scientist. Much of the effort in automating the process was in the provision of 'joining services' e.g. the "RETRIEVE" service described in Section 3. We aim to explore in future work, how many of these joining services are generic (such as reformatting) and how many must be custom made for each experiment.

Although we have been successful in developing and running workflows, we have also encountered some of the challenges inherent in unattended distributed computing. Reliance on public external services does open up the possibility of unforseen failure. The difficulty in tracing the cause of failure means the workflow enactment engine currently provides little support for recovering the experiment other than providing intermediate results found at that point and offering to repeat the experiment. Future development will try and increase this support. Consideration will also be given to issues of security. Although historically the academic scientist has not undertaken secure interactions with bioinformatics Web sites, the unattended operation of these experiments does raise some unease with users. There are also security issues with examination and comment upon data by third parties.

Our investigation of WBS shows workflows not only performing *in silico* experiments in a replicable, extensible manner, but with such ease and clarity of results they have become a primary tool in the authors quest to produce a fully comprehensive map of the WBSCR. The strength of biological results combined with automation not only vastly reduces the drudgery and complexity of the task in hand, but also reduces the potential of human error and increases productivity.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the assistance of the whole myGrid consortium. This work is supported by the UK e-Science programme EPSRC grant GR/R67743. M.T. and H.T. 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 *Proceedings of the UK e-Science All Hands Meeting 2003*, pp. 459–466.

Altschul,S.F., Thomas,L., Madden,A., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.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. *Brief. Bioinform.*, **5**, 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 Res.*, **9**, 428–436.

Ewart,A., Morris,C., Atkinson,D., Jin,W., Sternes,K., Spallone,P., Stock,A., Leppert,M. and Keating,M. (1993) Hemizygosity at the elastin locus in a development disorder, Williams syndrome. *Nat. Genet.*, **5**, 11–16.

Foster,I. and Kesselman,C. (eds) (2003) *Blueprint for a New Computing Infrastructure*, 2nd edn. Morgan Kaufmann Publishers, San Mateo, CA.

Foster,I., Kesselman,C., Nick,J. and Tuecke,S. (2002) The physiology of the Grid: an open Grid services architecture for distributed systems integration. Technical report of the Global Grid Forum.

Goble,C., Wroe,C., Stevens,R. and myGrid consortium (2003) The myGrid project: services, architecture and demonstrator. In *Proceedings of the UK e-Science programme All Hands Meeting*, University of Manchester, Manchester, UK.

Hendler,J. (2003) Science and the semantic web. *Science*, **299**, 520–521.

Hillier,L.W., Fulton,R.S., Fulton,L.A., Graves,T.A., Pepkin,K.H., Wagner-McPherson,C., Layman,D., Maas,J., Jaeger,S., Walker,R. *et al*. (2003) 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. *Trend Genet.*, **9**, 418–420.

Klyne,G. and Carroll,J.J. (2003) Resource description framework (RDF): Concepts and abstract syntax. *W3C Proposed Recommendation.*

Krishna,A., Tan,V., Lawley,R., Miles,S. and Moreau,L. (2003) myGrid Notification Service. In *Proceedings of the UK e-Science All Hands Meeting 2003*, Nottingham, UK.

Morris,C. (1988) The natural history of Williams syndrome: physical characteristics. *J. Paediatr.*, **113**, 318–326.

Osborne,L. (1999) Williams–Beuren syndrome: unraveling the mysteries of a microdeletion disorder. *Mol. Genet. Metab.*, **67**, 1–10.

Osborne,L., Li,M., Pober,B., Chitayat,D., Bodurtha,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. *Nat. Genet.*, **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. *Am. J. Human Genet.*, **66**, 47–68.

Preus,M. (1984) The Williams syndrome: objective definition and diagnosis. *Clin. Genet.*, **25**, 422–428.

Quan,D., Huynh,D. and Karger,D.R. (2003) Haystack: a platform for authoring end user semantic web applications. In Fensel,D., Sycara,K. and Mylopoulos,J. (eds), *Proceedings of the 2003 international semantic web conference (ISWC 2003)*, Lecture Notes in Computer Science 2870, Springer, pp. 738–753.

Senger,M., Rice,P. and Oinn,T. (2003) SoapLab—a unified Sesame door to analysis tools. In *Proceedings of the UK e-Science All Hands Meeting 2003*, Nottingham, UK.

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 Mol. Genet.*, **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. *Am. J. Human Genet.*, **64**, 118–125.

Valero,M., de Luis,O., Cruces,J. and Pérez 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.