

Exploring Williams-Beuren Syndrome Using myGrid

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

^a Department of Computer Science University of Manchester Oxford Road Manchester United Kingdom M13 9PL	^b University of Manchester Academic Unit of Medical Genetics St Mary's Hospital Hathersage Road United Kingdom M13 0JH
^c European Bioinformatics Institute Wellcome Trust Genome Campus Hinxton Cambridge United Kingdom CB10 1SD	

March 25, 2004

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, 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 with such a goal. 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 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.

As part of ongoing efforts to produce a complete map of the Williams-Beuren Syndrome region, 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 characterise 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 increasing 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 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.

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. ^{my}Grid 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. ^{my}Grid 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. ^{my}Grid not only supports the creation of the experimental protocol (the workflow), but also the management of the inputs, outputs, intermediates, hypotheses, findings and 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. In this paper we describe the *in silico* experiments required for exploring WBS, the use of the ^{my}Grid services to implement and manage the running of those experiments and their results. We show how ^{my}Grid has been successfully used to extend the genetic map into the WBSCR; find and characterise genes within this important region of human chromosome 7. In this work, ^{my}Grid has demonstrated the use of output from the UK e-Science programme to answer real questions within biology.

References

- Morris, C. (1988). The natural history of williams syndrome: physical characteristics. *Journal of Paediatrics*, 113:318–326.
- 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 ^{my}Grid. Accepted for publication at Intelligent Systems for Molecular Biology (ISMB) 2004.
- Stevens, R. D., Robinson, A. J., and Goble, C. A. (2003). ^{my}Grid: personalised bioinformatics on the information grid. *Bioinformatics*, 19:i302–i304.