Bioinformatic workflows: G-PIPE as an implementation

Alexander Garcia[1,2], Samuel Thoraval [1,3] , Leyla J. Garcia[4], Yi-Ping Phoebe Chen [2,5] and Mark A. Ragan[1,2]

[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Australia
[2] Australian Research Council (ARC) Centre in Bioinformatics
[3] Current address: Bioinformatique, Université Montpellier II, Montpellier, France
[4] Faculty of Information Technology, Fundacion Universitaria San Martin, Bogota, Columbia
[5] School of Information Technology, Deakin University, Burwood, Vic 3125, Australia
Correspondence to: Alexander Garcia (a.garcia@imb.uq.edu.au)

ABSTRACT
We present G-PIPE, a graphic pipeline generator for PISE that allows the definition of pipelines, parameterization of its component methods, and storage of metadata in XML formats. Our implementation goes beyond macro capacities currently in PISE. As the entire analysis protocol is defined in XML, a complete bioinformatic experiment (linked sets of methods, parameters and results) can be reproduced or shared among users. We also discuss the role of ontologies as as guidance systems in order to provide users with the possibility to define abstract work-flows, and execute them. A relevant baseline ontology is presented. Availability: http://if-web.imb.uq.edu.au

INTRODUCTION

Computational methods of problem solving need to interleave *information access* and *algorithm execution* in a problem-specific workflow. In complex domains like molecular biosciences, workflows usually involve iterative steps of querying, analysis and optimisation. Bioinformatic experiments are often workflows; they link analysis methods that typically accept an input file, compute a result, and present an output file. Query workflows are sometimes implemented over relational database management systems (Wong 2000; Haas *et al.* 2001) and in such cases can be built using SQL statements. Analysis workflows, on the other hand, provide a path to discover information beyond the capacities of simple query statements, but are much less easy to implement within a common environment.

Systems such as W2H (Ernst *et al.* 2003) and PISE (Letondal 2001) provide some tools that allow methods to be combined. W3H (Carver and Mullan 2002) is a task framework that allows integration of methods available under W2H. In the case of PISE, the user can either define a macro using BioPerl (www.bioperl.org), or use the interface provided and register the resulting macro. Macros cannot be exchanged between PISE and W2H although they provide GUIs for more or less the same set of methods (EMBOSS: Rice *et al.* 2000). Indeed, macros cannot be shared even among PISE users. G-PIPE provides a real capacity for users to share and define complete experiments (methods, parameters, and meta-information), substantially mitigating the syntactic complexity that this process involves. We have tested G-PIPE by defining different pipelines and exchanging results between different PISE servers. One of these pipelines was PATH (Del Val *et al.* 2002).

SYNTACTIC COMPONENTS AND WORKFLOW TERMINOLOGY

The workflow language presented here closely follows the concepts presented by Lei and Singh (1997) and Stevens *et al.* (2001). We have adapted these meta-models to bioinformatic analysis processes. We present key definitions below; a more-extensive presentation of these terms and concepts is in preparation:

- *Input data object:* a collection of input data.

- *Transformer:* the atomic work item in a workflow. In analysis workflows, it is an implementation of an analysis algorithm (analysis method).
- *Pipe component:* the entity that contains the required input-output relation (*e.g.* information about the previous and subsequent tasks); assures syntactic coherence.
- *Output object:* the result of the transformation applied to one or more input data objects.
- *Task:* a defined piece of work. One analysis method, together with its parameters and input data object(s), constitutes a task.
- *Stage:* an instance within a workflow, containing tasks, annotation, and output. A stage can contain more than one transformer.
- *Workflow*: a group of stages with interdependencies. It is a process bound to a particular resource that fulfills the process.
- *Parameters:* experimental conditions relevant to a particular transformer.
- *Annotation:* meta-information relevant to the experiment. Annotation is the main source of knowledge by which other researchers can understand and reproduce the experiment.
- *Protocol:* a set of information that describes an experiment. A protocol contains workflows, annotations, and information about the raw data.

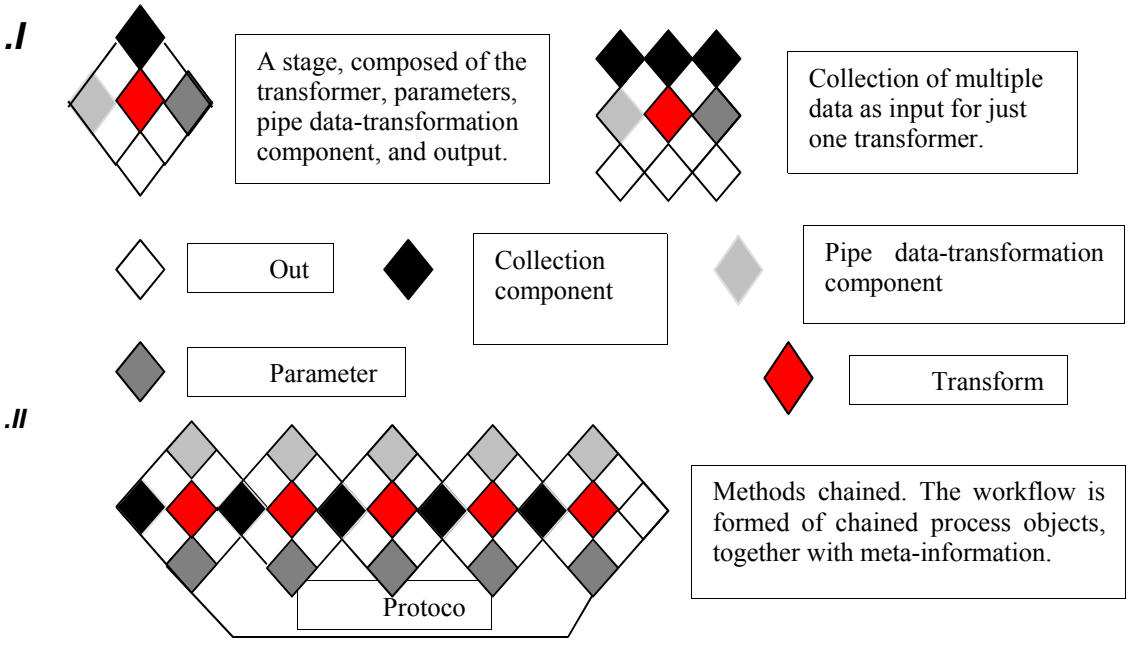

Fig. 1. Syntactic components describing bioinformatic analysis workflows

G-PIPE, A WORKFLOW GENERATOR

G-PIPE is a flexible workflow generator for PISE. We define each analysis protocol as an XML file; complete information, including metadata, is included. Our implementation is technically simple and reuses code from PISE. The information needed to generate the interfaces (HTML form and corresponding CGI) is contained in the PISE modules. The scripting capacities available in PISE come from a module that inherits the PiseApplication module from the PISE/Bioperl API.

The overall architecture of G-PIPE is shown in Fig. 2. A Java applet provides the user with an exploratory tool for browsing and displaying methods and protocols. The user interacts with the HTML forms to define a protocol. Synchronisation is maintained between client-side display and server-side storage using Javascript. Server-side persistency is maintained through serialised Perl objects that describe the experiment. The object is translated into two user-accessible files: an XML file to share and reload protocols, and a Perl script. A new lightweight PISE/Bioperl module, PiseWorkflow, lets workflows to be built and run atop PiseApplication instances. This module

supports independent branched tasks in parallel, and report errors and results into an HTML file.

G-PIPE allows a workflow to be defined from its atomic components. The user selects the methods, sets parameters, defines the chaining of different methods, and selects the server on which these will be executed. G-PIPE creates an XML file and a Perl script that describe the experiment. The user can monitor the status of workflow execution, and has access to intermediary results. A workflow built with G-PIPE can run its analyses on different, geographically disperse PISE servers.

DISCUSSION AND CONCLUSIONS

Integrating molecular biology information has syntactic and semantic components. It not limited to issues of databases, but must include analytical methods and results, as well as annotation and other metadata. A close relationship exists between query and analysis.

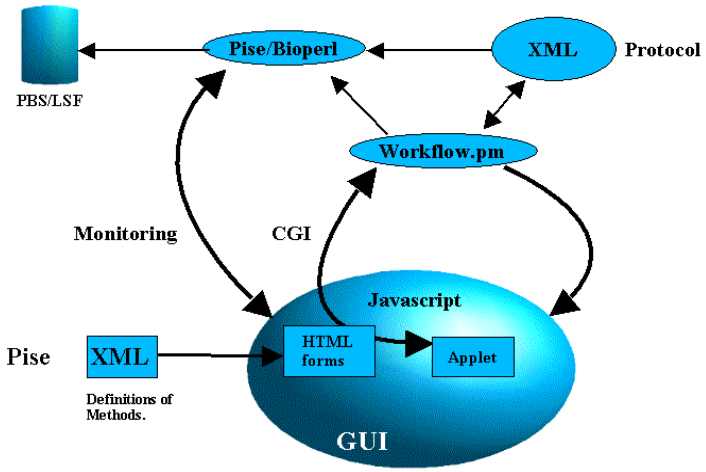

Fig. 2. Architecture of G-PIPE.

Our implementation extends the capabilities of PISE to allow the creation and sharing of customised, reusable and shareable analysis workflows. The same model is directly applicable to JEMBOSS and W2H, and would allow experiments to be exchanged across all three systems. G-PIPE has been designed for ready adaptation to newer versions of PISE. A robust workflow management system should ideally have a client/server architecture that extends beyond simple chain-of-steps capability. For example, agents should identify resources and schedule jobs across a computational grid. PISE servers can communicate with G-PIPE clients to coordinate the distribution of jobs.

The framework we have established is flexible and, in principle, extensible to queries as well: collections of data could easily come from *e.g.* a relational database, and query operations could be made an integral part of the protocol. Similarly, our framework does not depend on any specific GUI.

A more accurate semantic representation of service capabilities and interaction is needed to separate domain knowledge from operational knowledge, and allow workflows to be built from semantic components. A solution in which services are dynamically organized while shielding the user from operational issues would be highly desirable. The addition of an information retrieval system, and definition of an operational ontology that describes methods, will be part of future development over G-PIPE.

Ontologies may be used in order to guide users when building workflows; the relationship between an output and an input is mostly syntactic, and throughout this work has been treated as such. Different data types that take part in those analysis processes as well as those categories for all the analytical algorithms may be better organized as an ontology. Such an ontology may not only

be used for defining workflows from the concepts (abstract workflows) and map them in to concrete executable ones but it may also provide a standard which may make it possible to share workflows among different clients. As part of the second step within G-PIPE development we are precisely working on such ontology.

ACKNOWLEDGEMENTS

We thank Dr Lindsay Hood for valuable discussions. ST thanks Université Montpellier II for travel support. This work was supported by ARC DP0344488 and CE0348221.

REFERENCES

Carver,T.J. and Mullan,L.J. (2002) A new graphical user interface to EMBOSS. *Comp. Funct. Genomics*, 3, 75-78.

Del Val,C., Ernst,P., Bräuning,R., Glatting,K.-H. and Suhai,S. (2002) PATH: a task for the inference of phylogenies. *Bioinformatics*, 18, 646-647.

Ernst,P., Glatting,K-H and Shuai,S. (2003) A task framework for the web interface W2H. *Bioinformatics*, 19, 278-282.

Haas,L.M., Schwarz,P.M., Kodali,P., Kotlar,E., Rice,J.E. and Swope,W.C. (2001) DiscoveryLink: a system for integrated access to life sciences data sources. IBM Syst. J., 40, 489-511.

Lei,K. and Singh.M. (1997) A comparison of workflow metamodels. *In:* Workshop on behavioral modeling and design transformations: Issues and opportunities in conceptual modeling. ER'97, 6-7 November 1997, Los Angeles.

Letondal,C. (2001) A Web interface generator for molecular biology programs in Unix. *Bioinformatics*, 17, 73-82.

Rice,P., Longden,I. And Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., 16, 276-277.

Senger,M., Flores,T., Glatting,K.-H., Ernst,P., Hotz-Wagenblatt,A. and Suhai,S. (1998) W2H: WWW interface to the GCG sequence analysis package. *Bioinformatics*, 14, 452-457.

Wong,L. (2000) Kleisli, a functional query system. *J. Funct. Programming*, 10, 19-56.