An Information Management System for Collaboration Within Distributed Working Environment

(Systems Paper)

Maria Samsonova, Andrei Pisarev, Konstantin Kozlov, Ekaterina Poustelnikova, and Arthur Tkachenko

St.Petersburg State Polytechnical University, St.Petersburg, 195251 Russia

Abstract. Over a period of several years we apply the systems biology approach to investigate the dynamic regulatory mechanisms controlling the expression of segmentation genes in Drosophila embryo. Due to ongoing data acquisition, development of new processing and analysis methods, as well as modification and improvement of old ones serious problems arose with data and workflows management. Different geographical location of research groups poses additional difficulties. To solve these problems we have developed an information management system using multiagent and REST architectures. This system is easily extendable to deal with new data processing and analysis methods, flexible in specification and modification of these methods, scalable and supports distributed processing and analysis of data.

1 Introduction

Recently the introduction of high-throughput techniques as well as digital recording devices and computers lead to accumulation of large volumes of data. There are hundreds of resources and applications available to a biologist via "command line" applications, databases, flat files, web forms or graphical user interfaces. Publishing of data and providing services via the Internet has a long-lasting tradition in biology. Taking advantage of the broad-bandwidth Internet connections, researches are able to connect remotely to computers to share research data, tools and computing power.

Traditionally a biologist needs access to dozens of data types and services to plan her experiments and analyze results. To obtain such an information a researcher needs to navigate and download data from many computers, process, integrate and analyze the downloaded information manually or to use complex scripts to overcome incompatibilities. This is a very difficult and tedious task, as resources are widely distributed, highly heterogeneous, diverse and autonomous.

The increase in data, the rapid growth in a number of analysis tools and the range of knowledge needed to interpret and use them requires to develop methods for at least partial automation of data and services integration. Among obvious advantages of such an automation are reduction in a number of routine

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 204–215, 2006.
© Springer-Verlag Berlin Heidelberg 2006

queries for wet lab biologists, possibility to perform and repeat data analysis multiple times, reduction in research cost, as well as the transparency of code and algorithms.

Currently several service oriented architectures (SOA) are used to integrate heterogeneous resources. CORBA, RMI and DCOM are mainly applied to integrate Intranet applications. SOAs based on Web services concept use specific protocols to access services. Though several widely recognized implementations have been attempted [1,2,3,4] some aspects of these technologies impede their use for creation of a highly integrated global biological data space. These are developing and incomplete standards, introduction of several versions of standards with different policies in the field of patenting by different organizations (W3C, OASIS, Grid, etc.), insufficient solution of safety questions, necessity to modernize the already developed programs. Besides, as we will show in section 3.2, XML-representation of some data types (images, BLOBS, matrices) can decrease the performance of application.

Contrary to all the architectures described the REST (Representational State Transfer) architecture uses URIs to identify resources, and a small, globally defined set of remote HttpMethods to access and manipulate the state of those resources [5]. HTTP is the protocol by which resources are accessed. REST proponents argue that the HTTP's minimal method set and semantics, as well as its ability to extend this method set as required is sufficiently general to model any application domain.

This paper presents the application inspired by REST. It is designed to automate management and analysis of information generated by the consortium of laboratories in USA, Russia and Western Europe. Here we describe the prototype of this system and demonstrate real-life scenarios of data processing and analysis.

2 Materials and Methods

Over a period of several years the consortium of laboratories from St.Petersburg Polytechnical University, the Ioffe Physical-Technical Institute (Russia), Stony Brook University, Los Alamos National laboratory (USA) and University of Amsterdam (the Netherlands) investigates the dynamical regulatory mechanisms which control the expression of segmentation genes in Drosophila embryo [6,7]. To solve this problem the systems biology approach is applied, which encompasses the acquisition of data on a large scale, mathematical modelling and simulations.

Due to ongoing data acquisition, development of new processing and analysis methods, as well as modification and improvement of old ones, serious problems have been encountered with data storage and management of application programs. Additional problems are created by different geographical location of performance sites, as often users need data or programs kept in another laboratory. These problems complicate data analysis and processing and decrease the efficiency of work as a whole. To automate information management and