Developing Scientific Workflows to support Biomedical Applications

Nigel Ferguson

Abstract:

Within the biomedical sciences, advanced molecular biological techniques (e.g. high throughput proteomics, etc.) have enabled many elements of biological systems to be characterised at a rapid pace. Much of this data is now stored in various databases accessible via the Internet. This expanse of data has also resulted in the creation of many software prediction and modelling tools.

This “data landscape” presents a number of computational challenges. The tools and data are spread over abundant sites, new sites are frequently introduced, there are no common interfaces and the underlying data is constantly changing.

Often, a series of these tools and databases are linked together to form a computational pipeline. Until recently, the linkage of such tools would be achieved by joining these components via scripting languages such as Perl. Although many projects have been successfully based on this simple approach, it has many drawbacks, including: slow development time, underlying tools not architected for distributed environments, the need for a high degree of programming knowledge, the fact that complex applications are prohibitively difficult to construct and the issues associated with extensive component reuse.

Recently, the approach to creating these computational pipelines has been radically changed using graphical environments known as Workflow Management Systems (WMS).

This environment enables a user to connect graphically a number of components (e.g. prediction tools) together to solve a problem that each individual application could not solve. Currently, these workflows are limited to relatively straightforward applications. The WMS environment is ideally suited to dynamic applications as new components can be added with ease and pre-existing components can be readily reutilised.

The project aims to demonstrate how the dynamic qualities of a WMS can used on complex biological systems. An example of such a system is the prediction of epitopes within the adaptive immune response.