**Parameter Sweep Workflows for Modelling Carbohydrate Recognition**

1. **Introduction**

Carbohydrate recognition is a phenomenon critical to a number of biological functions in humans including highly specific responses of the immune system and the selective synthesis of functional proteins. Unlike polypeptides and proteins, oligosaccharides have been observed to have dynamic properties with extensive ability to occupy different conformations over time and space. Understanding the dynamic behaviour of oligosaccharides should provide clues towards the mechanisms which lead to specific and selective recognition of carbohydrates by proteins.

Computer programs which can provide insight into biological recognition processes have significant potential to contribute to biomedical research if the results of the simulation can prove consistent with the outcome of conventional wet laboratory experiments. Introduction and validation of these *in silico* tools would enable bio-scientists to focus their resources and better plan experiments by allowing them to visualise potential interactions and determine the best molecules to investigate in the wet laboratory. This not only reduces time and cost but also increases the numbers of molecules screened.

Unfortunately, there are several reasons why bio-molecular simulation packages are still not widely used among researchers. The simulations are usually very time-consuming where one simulation run can take weeks or months to complete on a single PC. Also, the command line interfaces provided by current simulation packages are far too complex for an average biologist. These tools do not provide support for automatic parameter sweeps or workflow type of execution either, that is often required by more complex scenarios. Finally, as a result of relatively low utilisation, these packages are not yet validated and tested to the required level.

The aim of our work is to provide solutions for all the above problems and to create a generic framework that can be utilised by bio-scientists to run massively parallel simulation workflows from a high level user-friendly environment. The solution allows bio-scientists with no Grid or parallel computing background to easily customise, parameterise, run and analyse complex simulation scenarios, and to provide useful feedback regarding the validation and refinement of *in silico* modelling tools.

In order to achieve these objectives, a generic high level user support and execution environment is required with the following main functionalities: (i) provide an easy to use preferably Web-based user interface for the non computer literate biologist end-user; (ii) provide an intuitive developer interface for selected computer trained biologists; (iii) support the creation of computational workflows orchestrating the execution of several components; (iv) support the creation of automatic parameter sweeps analysing several parallel scenarios based on different input parameters; (v) allow the mapping of workflow components to distributed computing resources (e.g. the UK National Grid Service); (vi) provide access to robust file storage systems and databases to manage input/output data; and finally, (vii) easily extendable by custom tools (for visualisation of results, for example) based on user demand and requirements.

There are several tools available that serve as good candidates for such framework, for example the MyExperiment community portal [1] or the P-GRADE Grid portal [2]. The latter has been selected in our project due to its support for parameter sweep workflow execution and its service status on the UK National Grid Service. However, the primary aim of the work was to identify the level of support and type of user environment that a biologist researcher with no or very limited computing knowledge requires to access and utilise existing e-infrastructures. Therefore, other existing tools and environments can also be customised and utilised to support this target user community.

2. **The User Scenario**

Modeling the interaction between receptors and ligands includes several steps where numerous scripts and simulations programs are utilized and the data is fed from one component to another. Working through such a complex scenario manually is a tedious process. Moreover, some of the steps require massive computing resources and the analysis of several parallel scenarios. Such user scenario is represented on figure 1. The process is divided
Parameter Sweep Workflows for Modelling Carbohydrate Recognition

into four separate phases where each phase uses several simulation and analysis programs and utilizes the output of previous stages. In phase 1 the pdb file describing the receptor is selected using public, local or user provided databases, checked, energy minimized and validated. Phase 2 represents similar steps in case of the ligand. In phase 3 the docking parameters and the target grid-space are defined first, followed by the actual docking simulations using the AutoDock package [3]. These simulations are run on a predefined number of parallel branches (typically around 10 parallel simulations are executed) each of which perform a genetic algorithm of a set number of evolutions (typically around 100). The best results of these simulations are then selected and energy minimization and molecular dynamics is performed, again in a number of parallel branches, on every selected docking result, using the MPI parallel version of Gromacs [4].

Figure 1 – User scenario for modeling receptor-ligand interaction

3. Workflow in the P-GRADE portal

The previously described user scenario has been implemented in the WS P-GRADE portal (the latest generation of the P-GRADE Grid portal) [2] and the execution of components has been mapped to resources of the UK National Grid Service. The WS-PGRADE portal supports the creation of computational workflows from a graphical user interface where workflow components can be mapped and submitted to Globus or g-Lite based Grid resources.Workflow components can be sequential or MPI parallel applications and any part of the workflow can be defined as a parameter sweep where the number of executions is determined by the number of input files.

The portal supports two different views. Workflow developers can construct the workflow graph and define the executables and input/output files associated with the different workflow components. This stage requires understanding of the P-GRADE developer environment. However, the complexity of the Grid and its middleware layer is completely hidden at this stage. The end user view allows scientist end-users to parameterize and submit the previously created workflows and to retrieve the results.

As the portal is based on the JSR-168 compliant GridSphere portal framework [5], it was fairly straightforward to extend it with custom portlets for visualization of the results. The visualization portlet uses the KiNG (Kinemage, Next Generation) [6] visualization tool and provides a convenient interface for end-users to compare and analyse the results of the simulations.

Figure 2 illustrates the P-GRADE workflow and its execution implementing the user scenario described in section 2. The figure shows workflow itself, the end-user interface where input command line and file parameters can be defined, and finally, part of the visualization interface to compare the results of two docking simulations.

4. Conclusions and future work

In order to increase the uptake of current e-Infrastructures new user communities with no Grid related knowledge have to be targeted. These researchers are reluctant to use this infrastructure due to its complexity and the very steep learning curve required. The aim of the above described work was to illustrate how high level Grid development and execution environments can be used to produce easy to use solutions with no or very minimal software engineering.
Parameter Sweep Workflows for Modelling Carbohydrate Recognition

These high level environments allow a small number of specifically trained researchers with more affinity towards software tools to serve the needs of larger user communities.

The implemented solution has been handed over and utilized by the target user community. Future work includes the creation of more customized workflow scenarios using the already developed and also new workflow components as building blocks. While the first stage workflow was created with significant support from the technical team, it is envisaged that future development will be based more on specialist end-user involvement.

Figure 2 – Workflow implementation, execution and visualisation in the P-GRADE Grid portal

References: