Drug Discovery & Development

Outsourcing Data Analysis via Secure Grid Technology

Grid technology is enabling companies to form short-term alliances for sharing expertise and data analysis tasks in a ‘virtual’ scientific ecosystem; provided the security of data can be assured, such technology presents great opportunities for improvement of the drug discovery process.

The pressures on the pharmaceutical industry from late stage attrition, generics and expiring patents are well documented. This is forcing a dramatic change in the way pharmaceutical companies think about sharing data and performing complex data analysis. Traditionally there has been a fortress mentality, which dictates that all data and analysis need to be done securely behind the company firewall, and any new software or approaches need to be installed and integrated into this environment only. However, there is increasing interest in outsourcing – not only for research to contract research organisations (CROs), but also for data analysis services. By moving to an ecosystem of partnering companies virtually sharing data, expertise and services, pharmaceutical organisations can improve the efficiency of drug discovery, enhance the biological context around targets and optimise drug effectiveness.

However, this is a significant change for pharmaceutical companies, as they are not set up to support this level of data interchange. Pharmaceutical organisations need to identify suitable providers, securely connect to outside services and incorporate these capabilities into flexible scientific applications. Of great interest and concern to many companies is the time and infrastructure required to set up new relationships via a virtual organisation (VO). A VO can be described as a finite collaboration between two or more parties – for example, a pharmaceutical company, a CRO to perform experiments and a group of data analysis providers. The duration of this relationship depends on the questions asked and the costs incurred by the interaction. Implementing such relationships has typically taken months and has not provided the needed flexibility and responsiveness to business issues. Alternatively, the use of Grid technology can reduce implementation time to weeks or even days, and enables companies to form short-term alliances to share expertise and data analysis tasks in a scientific ecosystem. Grid technology uses the resources of many separate computers connected by a network (usually the internet) to solve large-scale computation problems and improve communication between organisations.

To facilitate data sharing and collaboration to support product development, the European Union provided €11 million to the partners of the SIMDAT project in 2004 for a four-year project. The central requirement for the SIMDAT project was to provide an environment that allowed data exchange between companies in a controlled and secure way, and facilitated data analysis and mining using Grid technologies. The funding has been used to develop the capabilities needed to support virtual collaborations across multi-company/vendor relationships in a secure way. SIMDAT is a cross-industry endeavour, covering the automotive, aerospace and meteorological, as well as the pharmaceutical industry.

This article explains the technological requirements to enable virtual organisations in the pharmaceutical industry, and reviews the implementation of a SIMDAT pilot project at pharmaceutical leader GlaxoSmithKline (GSK) to deliver a protein annotation pipeline that combines information from multiple global GSK sites, commercial providers and academic centres of excellence.

INFRASTRUCTURE FOR VIRTUALISATION

Grid Enabling Components

Applications that communicate over the internet do so using standards protocols called Web Services, which define formats and capabilities that are supported by each computing service. Grid Services are a specialised kind of Web Services that support long-lived, computer-intensive activities, which may require the sharing of data between organisations and use of computing resources from multiple providers. The Grid Resources for
Industrial Applications (GRIA) product, from IT Innovations, is a protocol designed to support the core requirements for business applications that make use of grid services to perform intensive tasks. It is designed to support business-to-business (B2B) collaborations across organisational boundaries in a secure, interoperable and flexible manner.

Within the SIMDAT project, GRIA has been integrated into the InforSense data analytics environment to enable external data analysis and annotation services to be easily used in workflows with internal data and services. GRIA was deemed essential to the SIMDAT project because data analysis in each industry sector requires very intensive computer modelling and data analysis activities. With the GSK pilot project, GRIA ensures that all data communication between GSK sites and other partners is stable and fail-safe, and can recover from any data loss or interruptions to communication.

Secure Communication
For security-conscious pharmaceutical companies, the inclusion of enterprise strength security is paramount, particularly when their data will be transmitted outside their firewall. E2E Security from NEC Laboratories Europe’s IT Research Division provides internet security services including encryption, integrity protection, and authentication from end-to-end – even in the presence of intermediate systems like proxies or message queuing systems.

Building Data Analysis Applications in a Virtual Environment
In order to establish a virtual organisation, the IT infrastructure must support the construction of transient and ad hoc applications that combine internal and external programmatic services. It is important that these applications be quickly constructed and deployed to end-users to achieve rapid business benefits. The InforSense data analytics platform is a workflow-based system that allows analytical applications to be constructed visually without the need for coding; it is based on data manipulation and analysis components, which can be extended to incorporate new data sources and algorithms. This type of visual programming enables non-software developer individuals to build and deliver analytical applications. The applications that are constructed can be then deployed to a wide user-base, via interactive web pages, making best practice data analysis available to support decision-making across the enterprise.

As part of SIMDAT, the InforSense data analytics platform has been integrated with GRIA and the NEC E2E security mechanism to provide the Grid and security components required to support the SIMDAT application (see Figure 1). With these extensions, the InforSense platform is ideal to support the construction of applications in a virtual organisation because all internal and external services can be combined together in one environment.

IMPLEMENTING A VIRTUAL PROTEIN ANNOTATION PIPELINE AT GLAXOSMITHKLINE

Internal and External Services
Understanding the mechanisms of disease is vital to developing an effective new drug with minimal side effects. Every effort is needed to understand how
targeting a particular protein will affect the disease and the rest of the body. Target identification and validation groups use many experimental techniques to compare disease and non-disease groups to identify possible targets. This often results in a set of gene sequences that need to be further investigated and annotated using specific algorithms, predictive models and biological content. GSK had traditionally performed these operations from a central resource centre and it was decided that building a virtual Master Sequence Analysis Pipeline would be a challenging test of the SIMDAT technology. An InforSense workflow was developed to connect to internal services to retrieve sequence annotation, SNP data, antigenic, and expression information from GSK sites in Stevenage (UK), Research Triangle Park (North Carolina, USA) and Upper Merton (PA, USA). This demonstrated that internal services could be easily incorporated into a workflow paradigm. GRIA and E2E were then used to make accessible services from ULB (Antigenic), Informatica (Drugability) and EMBL (Drugability) (see Figure 2). While developments were required as part of the project, this scenario was successfully implemented and rigorously tested. This approach has proven that GSK can use its existing capabilities alongside external services and benefit from a wider knowledge base of expertise to better understand drug targets.

Lessons Learnt
The most important lesson learnt from the project is that it is possible to build a secure ecosystem of communicating companies and universities for the better understanding of scientific data. By combining the capabilities of the InforSense data analytics platform, GRIA and E2E, SIMDAT has illustrated that Grid technologies can be beneficial in multiple industries. The usability of the InforSense workflow environment has proven to be essential in being able to quickly drag and drop data and analysis tools together, and test a new system. This type of iteration and testing would have been very time-consuming using conventional programming approaches.

Future Directions
The trend towards drug discovery ecosystems of companies and universities is expected to grow as pharmaceutical companies start to break down the barriers around their organisations. As the drug industry becomes more like other industries – such as automotive and aerospace – it will need approaches such as those demonstrated in this SIMDAT project. Due to the complexity and security-aware nature of the marketplace, it is expected that the ecosystem will need to be regulated to trusted providers rather than actively seeking new services and using them immediately. The pharmaceutical industry is still a long way away from that level of openness. However, developments are needed to cope with the complex semantics found in the life sciences to enable companies and service providers to communicate effectively; it is hoped that the efforts of the Semantic Web will further support the opening up of collaboration.

The combination of internal and external service providers and CROs will require new types of business intelligence/analytics tools to monitor and manage the outsourcing of services and research. This complex problem provides significant challenges in communication and coordination across companies and services, but will present great opportunities for improvement of the drug discovery process.