THE SEEK: A PLATFORM FOR SHARING DATA AND MODELS IN SYSTEMS BIOLOGY

K. Wolstencroft[1], S. Owen[1], F. du Preez[2], O. Krebs[4], W. Mueller[4], C. Goble[1], and J. L .Snoep[2,3,*]

1. School of Computer Science, University of Manchester, UK

Katherine.wolstencroft@manchester.ac.uk

Stuart. owen@manchester.ac.uk

Carole.goble@manchester.ac.uk

Tel: +44 161 275 6195

Fax: +44 161 275 6236

2. Manchester Interdisciplinary Biocentre, University of Manchester, UK

franco.dupreez@gmail.com

3. Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

jls@sun.ac.za

Tel: +27218085844

Fax: +27218085863

4. Heidelberg Institute for Theoretical Studies (Hits), gGmbH, Germany

Wolfgang.Mueller@H-ITS.ORG

Olga.Krebs@ H-ITS.ORG

Tel: +49 (0)6221 - 533 - 231

Fax: +49 (0)6221 - 533 - 298

Running Title: The SEEK:Data and Models Sharing in Systems Biology

* CORRESPONDING AUTHOR AT 3.

Submitted to Methods in Enzymology.

Published as: doi:10.1016/B978-0-12-385118-5.00029-3

ABSTRACT

Systems biology research is typically performed by multidisciplinary groups of scientists, often in large consortia and in distributed locations. The data generated in these projects tends to be heterogeneous and often involves high-throughput omics analyses. Models are developed iteratively from data generated in the projects and from the literature. Consequently, there is a growing requirement for exchanging experimental data, mathematical models and scientific protocols between consortium members and a necessity to record and share the outcomes of experiments and the links between data and models. The overall output of a research consortium is also a valuable commodity in its own right. The research and associated data and models should eventually be available to the whole community for reuse and future analysis.

The SEEK is an open-source, web-based platform designed for the management and exchange of Systems Biology data and models. The SEEK was originally developed for the SysMO consortia (Systems Biology of MicroOrganisms), but the principles and objectives are applicable to any Systems Biology project. The SEEK provides an index of consortium resources and acts as gateway to other tools and services commonly used in the community. For example, the model simulation tool, JWS Online, has been integrated into the SEEK to enable model simulations and to allow new experimental data to be compared to simulation results, and a plugin to PubMed allows publications to be linked to supporting data and author profiles in SEEK.

The SEEK is a pragmatic solution to data management which encourages, but does not force, researchers to share and disseminate their data to community standard formats. It provides tools to assist with management and annotation as well as incentives and added value for following these recommendations. Data exchange and reuse rely on sufficient annotation, consistent metadata descriptions and the use of standard exchange formats for models, data and the experiments they are derived from.

In this chapter we present the SEEK platform, its functionalities and the methods employed for lowering the barriers to adoption of standard formats. As the production of biological data continues to grow in Systems Biology and in the Life Sciences in general, the need to record, manage and exploit this wealth of information in the future is increasing.

1. INTRODUCTION

The number of Systems Biology research projects has grown rapidly over the last decade. Some of these projects are very large, for instance SysMO (http://www.sysmo.net), a European project studying the Systems Biology of MicroOrganisms, consists of 320+ scientists working in more than 120 research groups, organised into 13 distributed projects across Europe. Typically such Systems Biology projects contain a heterogeneous group of scientists with a variety of life science, informatics and computational modelling backgrounds. In addition to heterogeneity in research background of the scientists, there can also be a great diversity between research projects, with large differences in data types, experimental procedures and models.

The multidisciplinary nature of Systems Biology projects necessitates a good exchange of data and models, such that an effective iterative cycle between experiment and model can take place. To make such an exchange possible it is necessary that the data and models are described according to certain standards, and that sufficient annotation and metadata is available. In this regard, data sharing in systems Biology faces the same issues as any data sharing in science. Reuse and future interpretation relies on common naming schemes and reporting standards and understanding the data in the context of the experiment(s) that created it. Conforming to these common standards, however, can be time-consuming and complicated, so the challenge for data management systems is to achieve this with minimal disruption to the daily activities of scientists by providing tooling, expertise and best practice guidelines.

Classic data management systems have focussed on prescriptive database and warehouse solutions for storing data. Such solutions are not always useful for the researchers however, as it would take a long time before the databases are developed and available. By that time, researchers may have large collections of unstructured legacy data. These solutions also require researchers to understand and adhere to rigid data structures and upload data in unfamiliar environments. For example, large scale scientific data sharing projects such as the BIRN (http://www.birncommunity.org/), caBIG (https://**cabig**.nci.nih.gov/*),* or GridPP (http://www.gridpp.ac.uk/), insist that each participant agrees to specific formats and model specifications and adapts to a common infrastructure. If data management resources have been budgeted for, the conversion of data to the prescribed standards is possible and such approaches can be successful, but in the general case, resources are limited and such solutions are too heavy weight for many consortia. In addition each individual must understand the standards and the data model in the new system in order to participate and must conform fully to this model.) The resulting data is uniform and of a high standard, but the time required for submission of data may result in low user participation with only small amounts of data being deposited.

An important aspect of data management is therefore a cost-benefit analysis. Here costs would not only be the development and maintenance of the infrastructure (software development and hardware) but would also include effort of researchers in the projects to make the data and models available. The benefit would be the availability and reusability of data and the availability of tools to work with the data. A good balance between costs and benefit must be found, and is not necessarily static. The greater the standardisation, the more reusable and comparable the data becomes, but there is a limit to the time and effort that can be expected from individual researchers without added benefits and incentives for their own work.

A more difficult aspect of data management is the reluctance of researchers to make their data available, especially before publication. Clearly, if data is only submitted to central repositories after publication the members of the consortia do not have full benefit from the available resources produced throughout the projects, which can hinder collaborations. Therefore, it is essential that control over sharing individual data items and models remains with the researchers and encourages incrementally sharing with colleagues and the wider community. In contrast, funding bodies are now making much clearer demands on researchers to share their results more quickly and many publicly funded initiatives must adhere to new data sharing policies. In SysMO, researchers are expected to pool their research capacities and know-how, and strongly promote the sharing of data, methods, models and results within the consortium and with the Systems Biology community.

To meet these data management challenges, technical as well as social, the SysMO-DB project has designed, developed and deployed a web-based infrastructure (the SEEK) and a methodology to overcome these barriers and enable sharing and exchange in systems biology. Although developed for the SysMO consortium, the SEEK platform addresses general issues in systems biology data sharing and is applicable and adaptable to other consortia. It is available as open source software and is designed for easy installation (http://www.sysmo-db.org/). The SEEK platform is consequently spreading. The Virtual Liver (http://www.sysmo-db.org/), EraSysBio+ (http://www.erasysbio.net) and UniCellSys (http://www.unicellsys.eu/) consortia are all examples of large Systems Biology networks that have adopted the SEEK.

In this chapter we describe the SEEK and illustrate its functionality with examples from the SysMO consortium. We start with an overview of the SEEK platform and an outline of its design principles. Next we discuss data management issues in more depth and show how the SEEK and associated tools assist scientists with the above mentioned problems. We finish the chapter with a more general discussion about the state of data sharing in the Life Sciences, and how suitable incentives can be found to encourage individuals and institutions to become more open.

2. THE SEEK PLATFORM

The SEEK is the name given to the whole SysMO-DB data sharing platform. Its development follows a rapid and incremental cycle with new functionality becoming available with each release (approximately every two months). As a result, the first version of the SEEK was deployed to the SysMO consortium within a year.

A rapid and user oriented development of the SEEK is ensured by frequent interactions in site-visits and workshops with a focus group of users, the SysMO-PALS. PALs (Project Area Liasons) are representatives from each SysMO project who are a mixture of experimentalists, modellers and informaticians. PALs are PhD students and Post-Docs, but not group leaders or project managers. This was a conscious decision to connect with people who could meet regularly and would be responsible for daily data generation and curation. The SysMO PALs are an extension of the SysMO-DB design team. New developments in SEEK are trialled with the PALs before release to the rest of the community and the PALs describe new requirements and request possible new features. For example, they recommended we provide a directory of SysMO consortium members, they highlighted the importance of SOPs and protocols in understanding experimental context, and they raised sensitive issues surrounding data security and access control. PALs also have extra responsibility in managing project membership and metadata. Consequently, the PALs gather intelligence from their projects and act as a dissemination mechanism for new developments with SysMO-DB.

2.1 Access to The SEEK

The SEEK web interface is the main user access point to the system and provides a gateway to all other SysMO-DB resources. The SEEK comprises of the SysMO Yellow Pages, an Assets Catalogue, a model simulation environment, and links to external resources. Figure 1 shows a screenshot of the SEEK, showing a summary of a collection of experiments and their associations. In addition to the web interface, the SEEK also has a REST interface, allowing programmatic access to resources and allowing future federations of SEEK instances from different communities. The following sections describe the different elements of the SEEK and how they interact.

2.2 The Yellow Pages

The Yellow Pages list the members of the consortium, their projects, institutes and expertise. This information helps foster links between SysMO projects and individual scientists, allowing people with the correct skill sets to be identified for collaborations. The Yellow Pages also links data and models and other SysMO assets to the scientists that produced them. Each asset is owned and controlled by the person registering it in SEEK. Individuals can decide whether to share immediately with the whole consortium, with their own project, or to restrict access to a few close collaborators.

2.3 The SEEK Assets Catalogue

The Assets catalogue is a registry of who owns what resources and where they can be found. SysMO assets include data, models, protocols, standard operating procedures (SOPs), workflows and publications. These assets may be held centrally in the SEEK, or they may be held in local project repositories. If they are held elsewhere, the SEEK indexes descriptions of the assets and can extract them from these external sources on demand, but it doesn't store them. Each asset (local or remote) is managed by the individual scientist who created and uploaded it. If assets are based on previously registered assets, an attribution system allows this to be recorded, which ensures scientists retain the credit for all their work. SysMO assets are registered with persistent URIs to allow stable referencing from publications.

Assets are associated with one another using the ISA hierarchy (Investigation, Study, Assay) (Sansone et al, 2008). ISA provides a framework for pooling data files and models in their experimental context. For example, data files can be associated with the SOPs used to create them, models can be associated with files containing construction data and validation data. Individual experiments (assays) and any associated assets can be grouped into larger studies and investigations, where the results of a combination of assays are required for biological interpretation.

ISA-TAB is being developed as a community standard and is a general tabular format for describing data from different types of omics experiments. By following such a community initiative, we enable future exchange of data with other public resources. In SEEK, we have extended the ISA omics concept to encompass mathematical models, to allow all SEEK assets to be described in the same ISA format. Figure 2 shows an ISA description of a set of SysMO experiments.

2.4 Access to External Resources

The SysMO-SEEK is a Gateway to other resources. SysMO users can analyse their data with commonly used tools from the community (for example, JWS Online (http://jjj.mib.ac.uk), a model simulation environment) (Olivier and Snoep, 2004]), or they can explore descriptions of asset in a community context (for example, using the BioPortal ontology repository) (Noy *et al*, 2009. By providing these services, we encourage uptake in the consortium and transform SysMO-SEEK from a static repository to an active, dynamic resource. Links to publications prepare the way for dissemination to the wider community. Direct links between the publications and associated data and models will allow the SEEK to become a supplementary material store for published work.

In the near future, the ability to run analyses through the SEEK will be implemented, driven by Taverna Workflows (Hull *et al,* 2006) (http://www.taverna.org.uk). A collection of workflows will assist in the meta-analysis of data registered in the SEEK.

Figure 3 shows the architecture of the SEEK, demonstrating how the different components interact. The SEEK adopts a modular approach, so access to other external resources can be added incrementally. The central piece of the architecture is the JERM (Just Enough Results Model), which allows the interpretation of relationships between assets and the understanding of their contents. The JERM is fully described in section 4.

The link to the JWS Online Simulation environment is also a crucial part of the SEEK architecture. It provides a suite of tools for model management, annotation and simulation. JWS Online can be accessed via SEEK, via a web browser or via web services. The interface gives access to the model parameters and initial conditions and allows the user to select between a number of functionalities such as time integration, steady state analysis, parameter scanning and metabolic control analysis. In addition, a reaction schema is linked to the model interface via which the user can display the rate equations, view the annotation and access external links to the models.

3. The Challenges of Data Management

The architecture of the SEEK platform allows for a flexible approach to uploading and linking assets. Such a record of data and models from a large research initiative is important in its own right, but the real challenge lies in being able to interpret the contents of the assets, which is necessary for comparison with other data sets and for further analysis.

In this section, we discuss the current issues with identifying and interpreting biological data and describe some community initiatives that are attempting to resolve and standardise data descriptions.

3.1 Biological Object Identity

Combining different types of biological data hinges on understanding exactly which biological objects interact and also on being able to identify the same biological objects in different datasets. Public data repositories often contain overlapping sets of information with the same biological objects having different names and identifiers. For example, Table 1 shows the different names for fructose bisphosphate aldolase A and different identifiers for this protein. This protein intersects several central metabolic pathways, including glycolysis, the pentose phosphate pathway, and the fructose and mannose metabolism pathway. Consequently, it features in many protein and pathway resources. It is therefore essential that we are able to map synonyms back to official names of genes and proteins to enable the integration of information and data relating to this biological object from multiple sources.

The need for consistency in naming biological objects and the use of unique identifiers is well recognised (Howe et al, 2008), but many scientists still use colloquial names to refer to genes, proteins and metabolites, for example, in daily practice. For their own use, this may not be problem, but for publishing results, or for querying across multiple data files, it can be impossible to determine if the same biological object is being referred to. It is also vital that biological objects are annotated with the most accurate description possible. For example, in the case of metabolites such as glucose, it may be necessary to define which isomer is being referred to. For example, if you know you are only measuring the concentration of D-glucose, you should annotate your data with the identifier for the D-glucose isomer. If you do not know if it is D-glucose, or a mixture of D and L-glucose, you should annotate your data with the identifier for glucose to avoid adding inaccurate annotation.

Public databases and commonly used resources provide a collection of ‘official’ names for biological entities. For example, UniProt IDs (Uniprot consortium, 2010) can be used to refer to specific proteins, or ChEBI IDs (Degtyarenko *et al*, 2008) can be used to refer to chemical entities. In SysMO-DB, we recommend the consistent use of these common vocabularies for all data values, and we have made such a list available for the most common data types in the SysMO projects. This allows identical biological objects to be identified and more easily compared across data sets.

3.2 Data in Context

Problems with keeping track of data extend beyond the use of biological identifiers. The flexibility and adaptability of biological systems to varying external conditions make the experimental context in which data are obtained of high importance. Data must be recorded with enough description to enable others to understand how it was generated and for what purpose. Descriptions about the data are often referred to as metadata (data about the data).

If, for example, the experimental protocol for preparing a particular biological sample is unknown, or the methods used for collecting particular measurements are not recorded, it is not possible to compare those results with others. If a model contains parameters that cannot be traced back to any source or to the conditions under which the parameters were determined, it is not possible to validate it, as model behaviour might be very dependent on these conditions.

Metadata is an important aspect of data management and data sharing. Annotating experimental results with a consistent set of information allows for easier discovery of relevant data as well as enabling others to potentially reuse it.

Metadata ranges from simple descriptions about who performed the experiment and when, to more detailed descriptions of growth conditions of biological species, sampling and preparation of samples, and description of the experimental conditions. All of this information is typically available if the work is published, but it is not computationally accessible. Also, data featured in published articles can be large, and stored externally in databases or supplementary data stores, and should ideally contain enough metadata for interpretation without external descriptions.

For many types of biological data, there are already community agreed standards for metadata reporting. These standards are often termed *minimum information models*. These models aim to describe the least required for others to interpret and reuse data in the future. The MIBBI portal (Minimum Information about Biological and Biomedical Investigations http://mibbi.org) (Taylor et al, 2008) is a collection of all current minimum information models in the Life Sciences. To date, there are over 30 in use in the community.

MIBBI models are a pragmatic solution to metadata collection. They recognise data annotation is a time-consuming and under-valued activity. By defining a minimum set of required metadata, they encourage more co-operation and buy-in from the community. In SysMO-DB, we also adopt the minimum information model idea with our *Just Enough Results Model* (JERM) (described fully in section 4).

As well as the MIBBI models, some communities also specify that metadata must be recorded using common vocabularies and ontologies. This makes querying the data computationally more straight-forward, but many laboratory scientists have no experience or expertise in using such resources. Vocabularies are numerous and can be complex and difficult to navigate. In SysMO-DB, we try and combat these problems by providing extra tooling to help understand which minimum information models, vocabularies and standards should be used in which circumstances.

Despite the provision of standards and vocabularies, data annotation is still time-consuming. It can add a huge overhead to the workload of the scientist, so extra incentives to stimulate an adequate annotation of data are being used. In fields such as transcriptomics, it is often a prerequisite to submit data to public repositories before publications are accepted. In other fields, journals are specifying requirements for supplementary data submissions.

In both of these cases, data is shared at the point of publication but not before. Scientists are reluctant to release their data until they are able to use it in a publication. Any data management system designed for scientists must respect this publication life-cycle and allow scientists to remain in control of data release and dissemination. In large consortia like SysMO, however, sharing within the consortium before publication is desirable. This sharing can be encouraged by making access control simple and ensuring consortium members are working under a common data sharing policy. Our aim is to provide good software tools to assist with annotation and to enhance the analysis capabilities (such as model simulations, integrative workflows, visualisation of results, versioning and recording of the scientific process), to stimulate rather than force scientists to annotate and upload their data.

As more data is produced in the public domain, the importance of making it available and reusable increases. There must, however, be clear incentives for scientists to describe and annotate their data sufficiently to make sharing possible and good practice. The software developed by SysMO-DB provides some of these incentives.

4. THE JERM INFRASTRUCTURE

The Just Enough Results Model (JERM) is the central organisational framework for the SEEK. It allows the exchange, interpretation and comparison between different types of data and results files. The JERM describes the minimum information required to identify and interpret assets. For example, for experimental data, the JERM describes what type of experiment was performed, who performed it, what was measured, and what the values in the datasets mean. It also allows for linkage between data, SOPs and models and therefore helps retain the context of the original experiment.

The JERM addresses the questions:

- What is the minimum information required to find the data?
- What is the minimum amount of information required to interpret the data?

The JERM follows the same principles as MIBBI. It is a minimum metadata specification to reduce the overheads of the scientists describing their data. Each asset has a title, a SEEK ID and an upload date. It is also associated with its creator(s) and a project. Other common elements help place the asset in context, for example, each asset should be associated with an assay and an assay type. If it is a data asset, it should be associated with SOPs, environmental conditions and factors studied. However, different types of data will require different JERMs at a more detailed level. The minimum information required to describe a microarray experiment, for example, is not the same as the minimum information required to describe a proteomics experiment using NMR. To make the data reusable for other scientist/studies it must be clear how the samples were prepared and what samples were used in the experiments.

In the SEEK, we promote the use of JERM compliance by providing JERM templates. The majority of SysMO scientists use Excel as a primary data management tool, so we provide JERM templates as spreadsheets to further encourage compliance and scientists upload their data in this format. It is also possible to acquire the same templates in alternative formats (e.g. XML schemas) for scientists using relational databases for storing local data.

4.1 JERM Harvesters and Extractors

The SysMO SEEK is an assets catalogue. It is a registry of assets stored in distributed project resources as well as assets stored centrally. In order to make use of all assets, wherever they are stored, the SEEK uses Harvesters for gathering data and extractors for interpreting their contents.

The retrieval and extraction of assets from the SysMO-SEEK is therefore a two stage process. Assets are registered in SEEK and searched over using their metadata. They are not retrieved from distributed project resources until required. If they are JERM compliant, further metadata can be indexed from within the asset using the SysMO Extractors.

When assets match search results, they are retrieved on demand, provided the user has permission to view and download them. The SEEK Harvesters control this process. They connect to a variety of local project resources, including wikis, content management systems, and relational databases. Harvesters can also be instructed to trawl over distributed resources at regular intervals in order to identify new SysMO assets automatically.

JERM compliance is optional. If data is uploaded in a JERM compliant format, querying over that data is easier, and more tools are available for using that data in analyses. Data can be uploaded in a non-compliant format, but there will be no attempt to parse or understand the contents. Adding incentives for data discovery and reuse encourages JERM compliance. This makes exchange and the eventual dissemination and export to other resources much more straight forward. It also helps with the registration of SysMO assets when they are stored in distributed project resources. JERM Harvesters and Extractors can be used to connect to these distributed assets on demand.

4.3 The SEEK and Data Management

Data in Systems Biology projects range from traditional molecular biology experiments through to the latest techniques in omics high throughput experiments. Typically, in these projects transcriptomics, proteomics and metabolomics analyses are conducted on the same samples, often alongside enzymatic activity analyses, protein-protein interaction studies and network analyses. Consequently, data can be large, complex, and in a variety of formats. The SEEK must cater for all these types of data, allowing storage (in the SEEK or at remote sites), and searches and comparisons between data sets. Consequently, JERM compliant templates for different types of experimental data are being produced in collaboration with SysMO researchers. These templates are potentially useful to other communities, so the collection will be made available as a SEEK resource.

4.4 The SEEK and Model Management

Model management in SEEK includes storage, annotation and simulation. Mathematical models play an important role in systems biology projects. They are crucial for understanding the behaviour of systems components, the description of experimental data, and the analysis and understanding of the systems under study.

Model management can be divided into a number of different tasks; model construction, simulation, validation, annotation, storage, and dissemination. Although model construction and validation would largely fall under the responsibility of the respective scientists, SysMO-DB provides tools to facilitate these steps, which would usually involve links with experimental data and models. In the case of a mathematical model that is constructed using a *bottom up* approach, such tools should enable visualisation of data sets used for the parameterization of the individual rate equations together with its goodness of fit. For model validation a different data set, for instance a time trace for model variables obtained on the complete system, could be used and then a visualization of the complete model together with the validation set should be possible. This example represents an idealized situation. There are not many models available in the scientific literature that show all data sets used for model construction and model validation. However, the SEEK provides the possibility to present large data sets along with the models and therefore promote these good modelling practices. Importantly such practices make the complete model building process transparent and reproducible. They would remove any uncertainties on how model parameters were derived.

In the SEEK, SBML (Hucka et al., 2003) is the recommended model format. Scientists are free to upload models in other formats, but the extra tools and functionality provided by JWS Online require SBML for use.

4.5 The SEEK and Process Management

Process management in the SEEK encompasses Standard Operating Procedures (SOPS) and protocols from laboratory investigations as well as data analysis protocols and model building methods. Conceptually, there should be no difference between these different types of protocol. Each describes the necessary conditions to understand and interpret the resulting data and each can be reused by other members of the consortium to perform the same experiments. In large, diverse consortia, where scientists are studying different organisms and different biological systems, the greatest added value can come from sharing methods and techniques rather than directly comparing data.

The multidisciplinary nature of systems biology projects means that cutting-edge technologies are often adapted and employed. Some require the development of new protocols. Sharing such protocols allows fast emergence of best practice and knowledge transfer between consortium partners. Unlike data, scientists are often willing to share protocols before their results are published, specifically if those protocols are obtained or modified from the literature.

As with data and models, standards for SOPs are recommended but SysMO-DB does not enforce them. Consortium members are free to upload or register SOPs in any format, but the *Nature Protocols* format is recommended (http://www.nature.com/nprot).

4.5.1 SOPs and Protocols

The distinction between SOPs and protocols is important in distributed projects. A Standard Operating Procedure is a protocol that has been agreed upon by a whole project or consortia. SOPs are essential for any part of the work that depends on standardising practices across the board. For example, when preparing cultures and samples that will be used in all subsequent experiments, it is important that they are prepared in exactly the same way to allow effective data comparisons. In SysMO, each project has a set of SOPs for the growth of cultures, which have typically been optimised over several iterations. For other experimental work, some protocols are used by the whole consortia, and some are only used by individuals.

4.5.2 Protocols for Informatics Experiments

In the SEEK, we make no distinction between laboratory experiments and informatics experiments. The bioinformatics analysis of data is simply considered to be another kind of experiment. Therefore, data and results should be recorded along with the SOPs and protocols used to produce it. In certain cases, however, the bioinformatics protocol may actually be executable. If the analysis was performed using a scientific workflow (for example in Taverna) (Hull *et al,* 2006), the workflow itself contains the protocol for the experiment and can therefore be shared and run again with the same data for verification, or with new data to perform related analyses. In the next phase of development, common data analysis tasks will be made available as Taverna workflows through the SEEK interface.

4.5.3 Protocols for Models

It is not yet common practice to write SOPs and protocols for modelling work, but capturing the process and the context of assumptions in the model is important, so we encourage the recording and storage of SOPs for modelling in SEEK. An important aspect of this work is identifying data that has been used for model construction and data that has been used for model validation.

4.6 Publications

The primary method for sharing scientific research remains the scientific publication. Publications can be registered in SEEK via a PubMed ID or a DOI. SEEK automatically matches author names to SEEK profiles and registers the publication abstract for searching. Any other asset can also be linked to a publication, which means that supplementary material for the paper can be associated directly from the SEEK.

5 The SEEK Functionalities: Annotating and Linking Assets

Annotation of assets, be it data or models is time-consuming and difficult. Scientists tend to start with annotation as and when they must do so for publication. For effective collaboration across distributed researchers, however, this practice has to be encouraged earlier.

For data annotation, the JERM templates provide a mechanism to help with this process. By using the JERM templates or schemas provided, SysMO scientists can produce JERM-compliant data. However, the templates only address the structure of the data. We must also consider the content. For mathematical models a MIRIAM annotation standard has been published (Le Novere et al., 2005), and we have implemented a tool in SEEK, OneStop to annotate models according to this standard and in the same time adhere to SBML (Hucka et al., 2003) and SBGN (Le Novere et al., 2009) model and network description standards as well.

In this section we introduce these tools, show how they are used in the SEEK and illustrate the strength of annotation in linking assets.

5.1 Data annotation and RightField

Typically, MIBBI standards dictate that particular values in a minimum information model should be annotated with terms from a particular domain-specific ontology. For example, when referring to the name of a chemical entity, that entity should be identified by its ChEBI entry (Chemical Entities of Biological Interest) (Degtyarenko et al, 2008), or when annotating SBML models, annotation terms should be taken from the SBO (Systems Biology Ontology) (http://www.ebi.ac.uk/sbo/). This is effectively another layer of annotation that is expected from SysMO scientists, but many are not familiar with the ontologies, or the advantages of uniform annotation for search and retrieval. Therefore, the approach we have adopted in SysMO-DB is to provide tools to make this process more accessible and straightforward.

RightField (Wolstencroft et al, 2011) is an open source application that provides a mechanism for embedding ontology annotation support in Excel spreadsheets. Individual cells, columns, or rows in spreadsheets can be restricted to particular ranges of allowed classes or instances from chosen ontologies. Bioinformaticians, with experience in ontologies and data annotation, can prepare RightField-enabled spreadsheets with embedded ontology term selection support for distribution across the consortium.

RightField supports the loading of ontologies (in OWL, OBO, or RDF format) (http://www.w3.org/standards/) directly from the BioPortal ontology repository, or from a local machine. When a spreadsheet has been marked-up with terms from selected ontologies, they are embedded into the Excel. Once marked-up and saved, the RightField-enabled spreadsheet contains embedded worksheets with information concerning the origins and versions of ontologies used in the annotation. This encapsulation stage is crucial. With everything embedded in the spreadsheet, scientists do not require any new applications to use it and they can complete annotation offline should they wish. This also makes the spreadsheets readily exchangeable and enables a series of experiments to be annotated with the same versions of the same ontologies even if the live ontologies change during this time. Ontology versions can be updated if the spreadsheet is opened again in RightField, but it is not automatic.

The RightField-enabled spreadsheet presents selected ontology terms to the users as a simple drop-down list, enabling scientists to consistently annotate their data without the need to understand the numerous metadata standards and ontologies available to them. The result is semantic annotation by stealth. RightField facilitates an annotation process that is quicker, less error-prone and more efficient. By combining JERM templates and embedded ontology terms with RightField, we provide an infrastructure that promotes and encourages compliance and standardisation. The result is a collection of data files with consistent annotation that is consequently easier to search and compare. Examples of these can be downloaded from the RightField website (http://www.rightfield.org.uk).

Figure 5 shows RightField being used to mark-up a transcriptomics data template with terms from the MGED ontology.

5.2 Tools for Model Annotation

The recommended standard for exchanging systems biology models is SBML, but SBML alone is not sufficient for a comprehensive understanding of the model. In Systems Biology, models are often used to simulate a specific system and contain variables and parameters which represent physical biological entities. Annotating the model with unique identifiers for molecular species (e.g. ChEBI), reaction steps (e.g. KEGG), and enzyme species (e.g. EC numbers), for example, allows model simulation results to be analysed in the context of experimental data and enables others to interpret the model more effectively. MIRIAM annotation can be used for this contextual understanding.

For small models, a standardized model description format might not appear to be that important. For example, the formulation of Ordinary Differential Equations (ODEs) with parameter values and initial conditions seems simple enough. However, when screening the scientific literature it quickly becomes evident that few models are described in sufficient detail that they can be reconstructed and simulations be repeated. This might reflect the aims of the scientists to illustrate a principle more than to build a realistic model, but still it is disconcerting that most of these models can never be used again. The JWS Online database was created to address this issue. It is both a curated models repository and a simulation environment. The OneStop tool is an extension to the JWS Online environment to allow MIRIAM annotation and the construction and editing of models.

OneStop provides an interface where users can define their model in a number of text fields in a web-browser. Subsequently, the model can be simulated using the JWS interface. Models can be defined from scratch, but the user can also upload SBML files or any of the models from the JWS Online or Biomodels databases (Le Novere et al., 2006). Models can be saved in SBML format and an automated SBGN schema generator and a tool for MIRIAM annotation are available. The annotation tool makes use of webservices from semanticSBML (Krause et. al. 2009).

Examples of text fields for model description are shown in Figure 6 and the model annotation field is shown in Figure 7. These tools are integrated into the SEEK environment

5.3 Linking Data and Models

Linking data and models relies crucially on the annotation of both the data sets and the model components. There is currently no community standard or fixed structure for expressing the connections between them. In SysMO-DB, we have been working on a number of scenarios for how data and models could be linked. At a basic level, SEEK users can specify that a particular dataset was used either in the construction or validation of a model. If data is the result of a model simulation run, we can also draw this distinction. However, much greater detail is needed for comprehensive integration.

For a number of metabolic models we have illustrated how data can be linked to the individual processes. For instance, for *bottom-up* models, a user could have an experimental data set for each of the model processes and on the basis of the data the user would formulate a mathematical equation.

In JWS Online it is possible to work with an isolated rate equation. Users can plot the rate equation with the parameter values used in the model (and he/she can change these parameter values). In addition, experimental data sets can be uploaded (for instance as excel files), and plotted together with the rate equation used in the model. In SEEK, we are developing mechanisms for easily importing/exporting data for plotting against models (for an example see Figure 8). Typically such data sets would be used for model construction. For model validation one could use data sets obtained with the complete organism/pathway, and such data sets would be linked to complete models.

7. INCENTIVES FOR SHARING DATA

The SEEK is a sharing initiative driven by funding councils in Europe, as a platform to assist the SysMO consortia members but also to ensure that the ever-increasing amounts of scientific data generated by public funding are made available to the community for further analysis and reuse. The SEEK provides a repository for all data and models from one funding initiative, creating a central focus for scientists involved in the initiative as well as a record of the research developed from it. It allows researchers to search and compare results or experimental techniques and include data from earlier work in wider studies. These outcomes are of benefit to the wider Systems Biology community, but there must be incentives for the SysMO scientists to spend time on data curation, annotation and sharing. The SEEK encourages participation by providing such incentives, which are; the provision of a safe haven for data and other assets, a set of tools for further analysis of these assets and for easy implementation of data and model standards in annotation, the opportunity for individuals to receive credit and attribution for their data contributions, and the ability to easily export assets to other public repositories. The following section describes these advantages for SysMO consortium members to adopt the SEEK.

7.1 Secure and Continuous Storage

Consortium members are obliged to make SysMO assets available to the community for 10 years as a condition of funding. If scientists opt to retain assets locally, the responsibility of ensuring they are available for others remains with them. During the day-to-day running of a project, this is often the case. However, when projects finish the individuals responsible for local upkeep and maintenance may move to new institutions.

If all SysMO assets are not uploaded to SEEK at the end of a project, the responsibility to make the data available long-term also remains with the scientists. SysMO-DB provides an archiving service to allow SysMO projects to publish all assets centrally at the end of their funding period, providing a guaranteed safe-haven for assets for an initial period of 10 years. This releases scientists from the overheads associated with maintaining individual resources and enables the whole consortium and others to benefit from the pooling of SysMO assets.

7.2 Credit and Attribution

Biological data can take months to collect and longer to analyse and publish. Traditionally, this data has only been used as evidence in the resulting publication, but data reuse is becoming more common as a result of large-scale analyses and the emergence of public repositories.

If data is adequately annotated and documented, it is potentially useful for future experiments and some data sets can even become widespread "reference" sets that are reused in multiple investigations.

In the Life Sciences, scientists are credited for their publications, but not traditionally for the actual data. Obtaining data of a good quality that can be used in multiple analyses is an advantage to the whole community. Therefore, the concept of data citation is becoming more popular (Nature Genetics Editorial, 2009) and mechanisms to enable this are now being proposed (http://www.datacite.org).

Ensuring SysMO scientists gain maximum credit for their work is an important incentive for registering and sharing. SysMO assets are associated with the profiles of their creators and registered with a unique and persistent URL to allow direct links to be made from publications and other online sources.

Attribution is an equally important issue. Experiments are often based on other experiments. SOPs are modified to improve efficiency, for example, or raw data is normalised or analysed. In these cases, the same scientist may not be responsible for the original and subsequent work, so it must be made clear which parts belong to which individuals. In the SEEK, credit and attribution are clearly visible. It is possible to credit other people for any work being shared and any asset can be attributed to any other, to signify that it was based on earlier work.

7.3 Export to Public Repositories

The SEEK is a unified interface to the outcomes of SysMO, but many journals require data to be submitted to public repositories before papers can be published. This is particularly true with Omics data. For example, microarray data must be submitted to ArrayExpress (Parkinson et al, 2005) or GEO (Barrett et al, 2009) before any paper is published relating to it. Such public repositories require data in particular formats to comply with community metadata standards. In SEEK, we plan to offer conversion services to allow one-click export, either by making use of tools from the ISA Infrastructure (Rocca-Serra et al, 2010) (in this case, the ISA Converter) (http://isatab.sourceforge.net/converter.html), or by directly mapping from the SysMO-JERM models.

The advantage for SysMO users is that data annotation and formatting only needs to be done once, at the initial registration with SEEK. In addition, we provide tools that makes adhering to such standards easier, for instance for mathematical models we have the OneStop tool for generating MIRIAM annotated SBML models, together with networks schema drawn according to SBGN standards.

8 The SEEK: Experiences

Since the initial release of the SEEK in SysMO, we have seen a gradual rise in uptake and use. There are already over 1700 assets registered in the SEEK and over 200 active users. As expected, we see a spectrum of compliance levels with registered assets. Some are registered with sparse metadata and remain unchanged, whilst others are richly described, or have incremental metadata additions to conform to the JERM.

We have, however, observed a much lower uptake of recommended formats and standards than we expected. For example, for models, SysMO-DB recommends SBML (which is also the community standard), but many in the consortium do not use it for the following reasons:

- It is not seen as fit for purpose
- It is still under constant development, and therefore is viewed as too unstable
- A lack of specific tooling support means it is difficult to import and export from applications already in use

These issues can, to some extent, be surmounted by simple interventions once they have been identified. For example, more tooling can be provided to help with format exchange from common applications (as in OneStop, the JWS Online Model Constructor), and the consortium can officially propose new directions to the standards developers to address shortcomings in the specification. In the meantime, the “Just Enough” principles of SysMO-DB ensure that consortium members are already free to share in any format until these new developments can be implemented.

The “Just Enough” design in SysMO-DB is the most fundamental part of the System. It is essential to provide a flexible model which users are free to interact with at different levels of compliance and detail, and at different times.

The flexibility of using a minimum model like the JERM encourages uptake and encourages social connections between consortium members. For assets that are poorly annotated, discussing their meaning and contents with the creator is the most efficient route to a contextual understanding. Therefore, data sharing and integration can be achieved through automated methods with the JERM extractors and harvesters, or through dialogue between consortium members with the SysMO-SEEK Yellow Pages. The social connections also tie individuals’ reputations to their assets. This encourages the addition of more metadata to prevent misinterpretation.

The next steps for SysMO-DB involve a greater focus on data analysis. Data exchange and sharing is becoming more popular, but the primary concern has been to encourage this behaviour and ensuring assets are recorded and archived. Now we have a growing collection of data and models, we need to provide more sophisticated ways of exploring and comparing them.

The overall SysMO-DB design methodology has been successful because we have focused on the specific concerns of the user community and built a solution that fits in with existing practices. Everything is designed in consultation with the SysMO PALs focus group, so they can help us identify bottlenecks and essential new features. Within the consortia, the PALs have formed their own network of young Systems Biology researchers with experience in data management and close collaborations between modellers and experimentalists.

The emergence of large-scale scientific consortia, in Systems Biology and other areas, coupled with the rapid development of more high-throughput experimental techniques, is driving changes in the way we record, reuse and reward data. Data management is consequently becoming more complex both locally and at a community level. To properly pool research outcomes and promote reuse, it must be easier for scientists to manage and publish data. This means providing tools for data storage and for data standardisation and analysis. The SysMO-DB project offers a suite of tools for a pragmatic data management solution to allow sharing in a large consortium and beyond with minimum impact on the daily work of researchers.

9 Acknowledgements

Funding: This work was funded by the BBSRC and the BMBF. SysMO-DB: Supporting Data Access and Integration.

10 References

Barrett, T., D. B. Troup, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res 37(Database issue): D885-90.

Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet 29(4): 365-71.

Degtyarenko, K., P. de Matos, et al. (2008). "ChEBI: a database and ontology for chemical entities of biological interest." Nucleic Acids Res 36(Database issue): D344-50.

Editorial. "Data producers deserve citation credit." (2009) Nat Genet 41(10): 1045.

Hull, D., K. Wolstencroft, et al. (2006). "Taverna: a tool for building and running workflows of services." Nucleic Acids Res 34(Web Server issue): W729-32.

Howe, D., M. Costanzo, et al. (2008). "Big data: The future of biocuration." Nature 455(7209): 47-50.

Hucka M, **Finney A**, **Sauro HM**, **Bolouri H**, **Doyle JC**, **Kitano H**, **Arkin AP**, **Bornstein BJ**, **Bray D**, **Cornish-Bowden A**, **Cuellar AA**, **Dronov S**, **Gilles ED**, **Ginkel M**, **Gor V**, **Goryanin II**, **Hedley WJ**, **Hodgman TC**, **Hofmeyr JH**, **Hunter PJ**, **Juty NS**, **Kasberger JL**, **Kremling A**, **Kummer U**, **Le Novère N**, **Loew LM**, **Lucio D**, **Mendes P**, **Minch E**, **Mjolsness ED**, **Nakayama Y**, **Nelson MR**, **Nielsen PF**, **Sakurada T**, **Schaff JC**, **Shapiro BE**, **Shimizu TS**, **Spence HD**, **Stelling J**, **Takahashi K**, **Tomita M**, **Wagner J**, **Wang J**; **SBML Forum**. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. **Bioinformatics.** 2003 Mar 1;19(4):524-31.

Le Novere, N. Le, Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., and Wanner, B.L. (2005) Minimal information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology 23:1509-1515.

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E.,Wegner, K., Aladjem, M., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Vill'eger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009) The Systems Biology Graphical Notation. Nature Biotechnology 27: 735-41.

Le Novère N., Bornstein B., Broicher A., Courtot M., Donizelli M., Dharuri H., Li L., Sauro H., Schilstra M., Shapiro B., Snoep J.L., Hucka M. (2006) BioModels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems *Nucleic Acids Res.*, 34: D689-D691.

Noy, N. F., N. H. Shah, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic Acids Res 37(Web Server issue): W170-3.

Parkinson, H., U. Sarkans, et al. (2005). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res 33(Database issue): D553-5.

Olivier, B.G and Snoep, J.L. (2004). Web-based kinetic modelling using JWS Online. Bioinformatics 20:2143-2144

Orchard S., Salwinski L., Kerrien S., Montecchi-Palazzi L., Oesterheld M., Stümpflen V., Ceol A., Chatr-Aryamontri A., Armstrong J., Woollard P., Salama J.J., Moore S., Wojcik J., Bader,G.D., Vidal M., Cusick M.E., Gerstein M., Gavin A.C., Superti-Furga G., Greenblatt J., Bader J., Uetz P., Tyers M., Legrain P., Fields S., Mulder N., Gilson M., Niepmann M., Burgoon L., Rivas J. de L., Prieto C., Perreau V.M., Hogue C., Mewes H.W., Apweiler R., Xenarios I., Eisenberg D., Cesareni G., Hermjakob H. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx).Nature Biotechnology 25: 894-898.

Rocca-Serra, P., M. Brandizi, et al. "ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level." Bioinformatics 26(18): 2354-6.

Sansone, S. A., P. Rocca-Serra, et al. (2008). "The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?"" Omics 12(2): 143-9.

Taylor, C. F., D. Field, et al. (2008). "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." Nat Biotechnol 26(8): 889-96.

Taylor C.F., Paton N.W., Lilley K.S., Binz P.A., Julian R.K., Jones A.R., Zhu W., Apweiler R., Aebersold R., Deutsch E.W., Dunn M.J., Heck A.J., Leitner A., Macht M., Mann M., Martens L., Neubert T.A., Patterson S.D., Ping P., Seymour S.L., Souda P., Tsugita A., Vandekerckhove J., Vondriska T.M., Whitelegge J.P., Wilkins M.R., Xenarios I., Yates J.R., Hermjakob H. (2007) The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology 25: 887-893.

The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010,
Nucleic Acids Res. 2010 January; 38(Database issue): D142–D148.

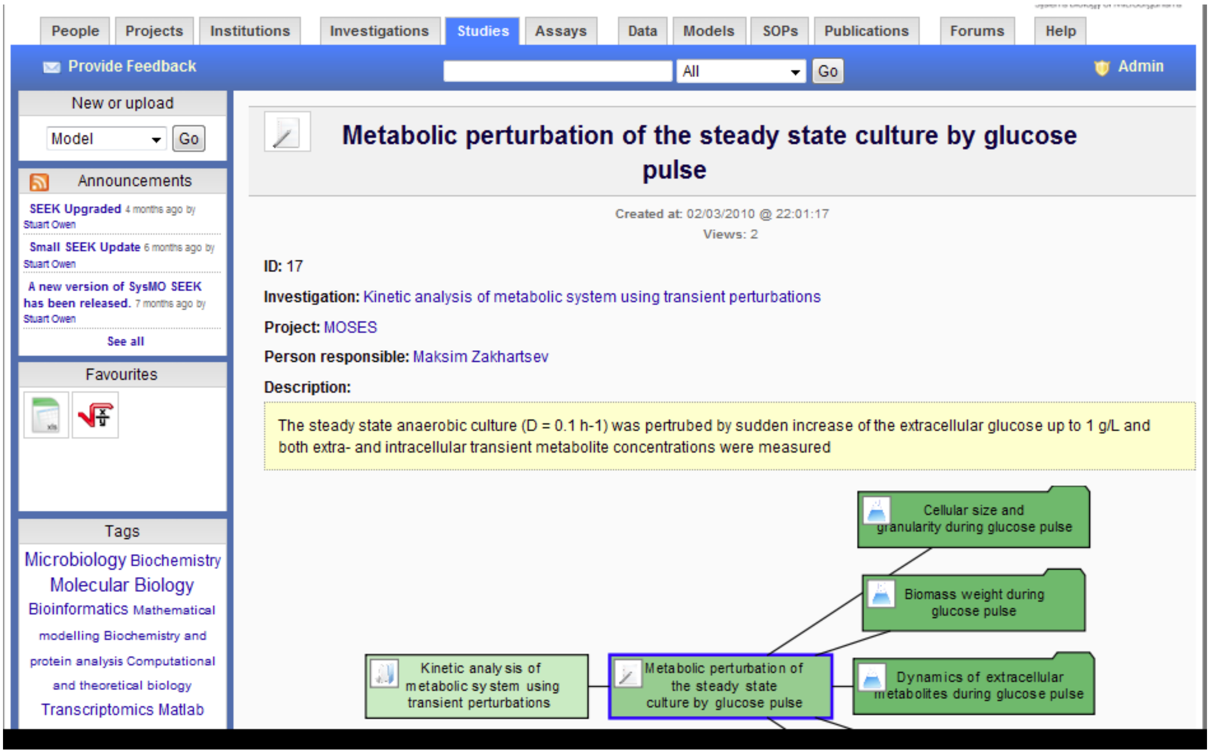
Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, et al. 2011 Minimum Information About a Simulation Experiment (MIASE). PLoS Comput Biol 7(4)

Whetzel, P. L., H. Parkinson, et al. (2006). "The MGED Ontology: a resource for semantics-based description of microarray experiments." Bioinformatics 22(7): 866-73.

Wolstencroft, K et al, (2011) RightField: Embedding ontology annotation in spreadsheets, Bioinformatics, In Press

Figure Legends:

Figure 1: A screenshot of the SysMO SEEK Interface


People
Projects
Institutions
Investigations
Studies
Assays
Data
Models
SOPs
Publications
Forums
Help
Provide Feedback
All
Go
Admin
New or upload
Model
Go
Announcements
Favourites
Tags
Kinetic analysis of metabolic system using transient perturbations
Metabolic perturbation of the steady state culture by glucose pulse
Cellular size and granularity during glucose pulse
Biomass weight during glucose pulse
Dynamics of extracellular metabolites during glucose pulse

SEEK Upgraded 4 months ago by Stuart Owen

Small SEEK Update 6 months ago by Stuart Owen

A new version of SysMO SEEK has been released. 7 months ago by Stuart Owen

See all

Microbiology Biochemistry Molecular Biology Bioinformatics Mathematical modelling Biochemistry and protein analysis Computational and theoretical biology Transcriptomics Matlab

Metabolic perturbation of the steady state culture by glucose pulse

Created at: 02/03/2010 @ 22:01:17

Views: 2

ID: 17

Investigation: Kinetic analysis of metabolic system using transient perturbations

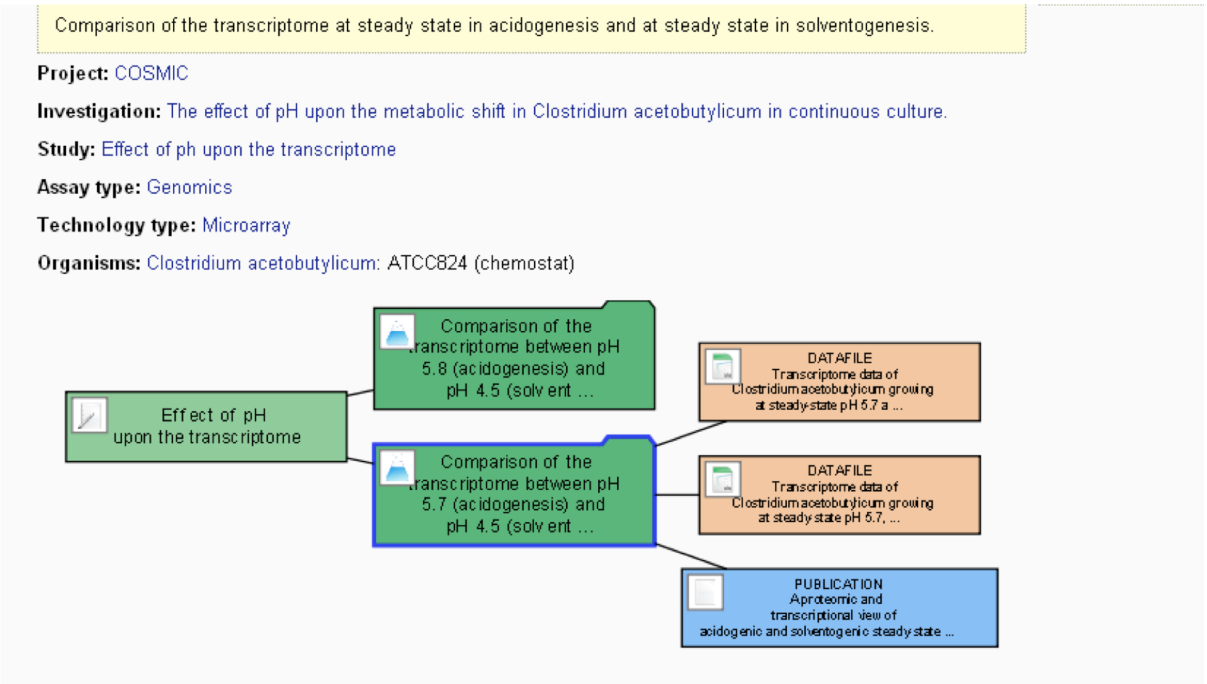
Project: MOSES

Person responsible: Maksim Zakhartsev

Description:

The steady state anaerobic culture (D = 0.1 h-1) was pertrubed by sudden increase of the extracellular glucose up to 1 g/L and both extra- and intracellular transient metabolite concentrations were measured

Figure 2: A screenshot of the ISA structure in SEEK and the interconnection of data and other assets in context of the experiments that created them.



Comparison of the transcriptome at steady state in acidogenesis and at steady state in solventogenesis.

Project: COSMIC

Investigation: The effect of pH upon the metabolic shift in Clostridium acetobutylicum in continuous culture.

Study: Effect of ph upon the transcriptome

Assay type: Genomics

Technology type: Microarray

Organisms: Clostridium acetobutylicum: ATCC824 (chemostat)

Effect of pH upon the transcriptome
Comparison of the transcriptome between pH 5.8 (acidogenesis) and pH 4.5 (solvent ...
Comparison of the transcriptome between pH 5.7 (acidogenesis) and pH 4.5 (solvent ...
DATAFILE Transcriptome data of Clostridium acetobutylicum growing at steady-state pH 5.7 a ...
DATAFILE Transcriptome data of Clostridium acetobutylicum growing at steady state pH 5.7, ...
PUBLICATION A proteomic and transcriptional view of acidogenic and solventogenic steady state ...

Figure 3: The architecture of the SEEK

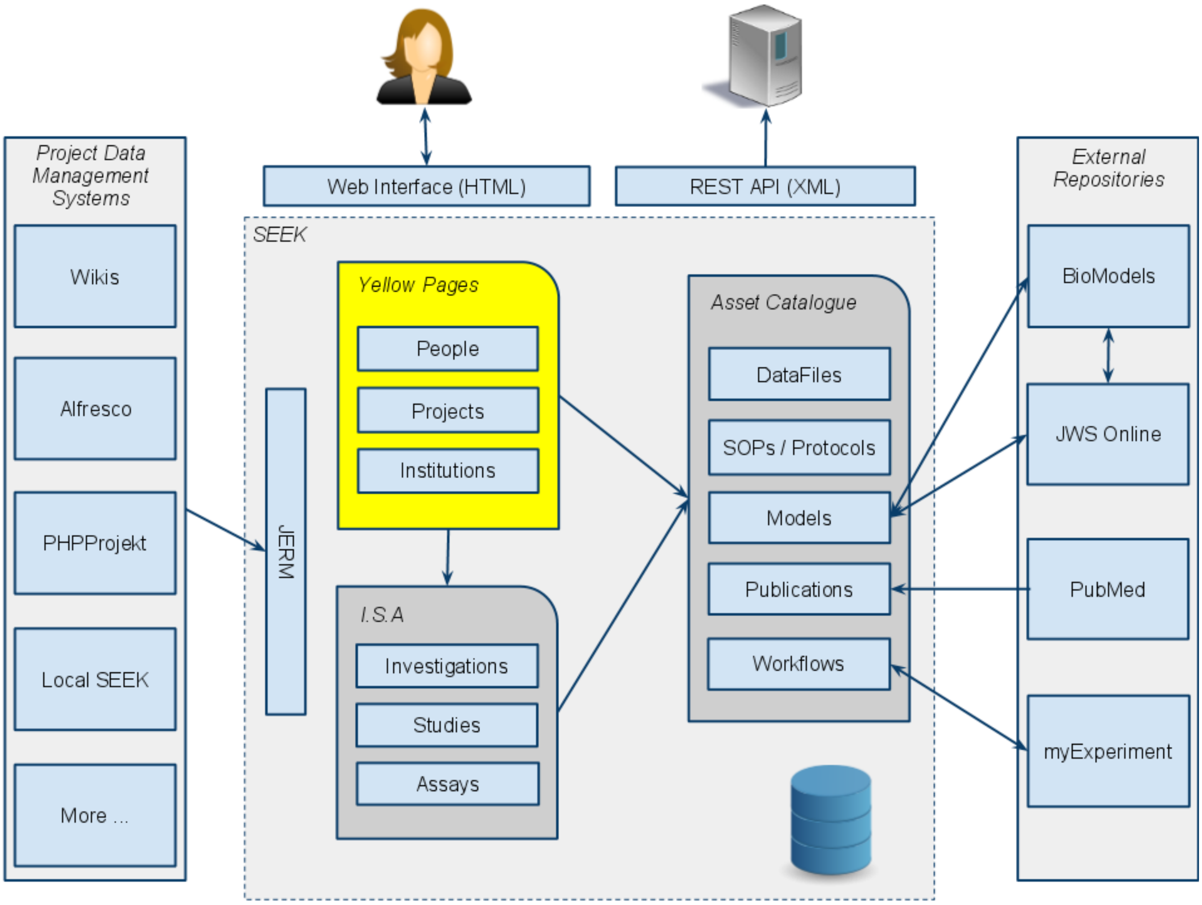


Figure 4: JERM Harvesters and Extractors are used in combination to retrieve SysMO assets held in distributed locations. The data is normally returned to the SEEK interface via the HTML pages, but can also be returned via REST services to enable computational access to the data for analyses.

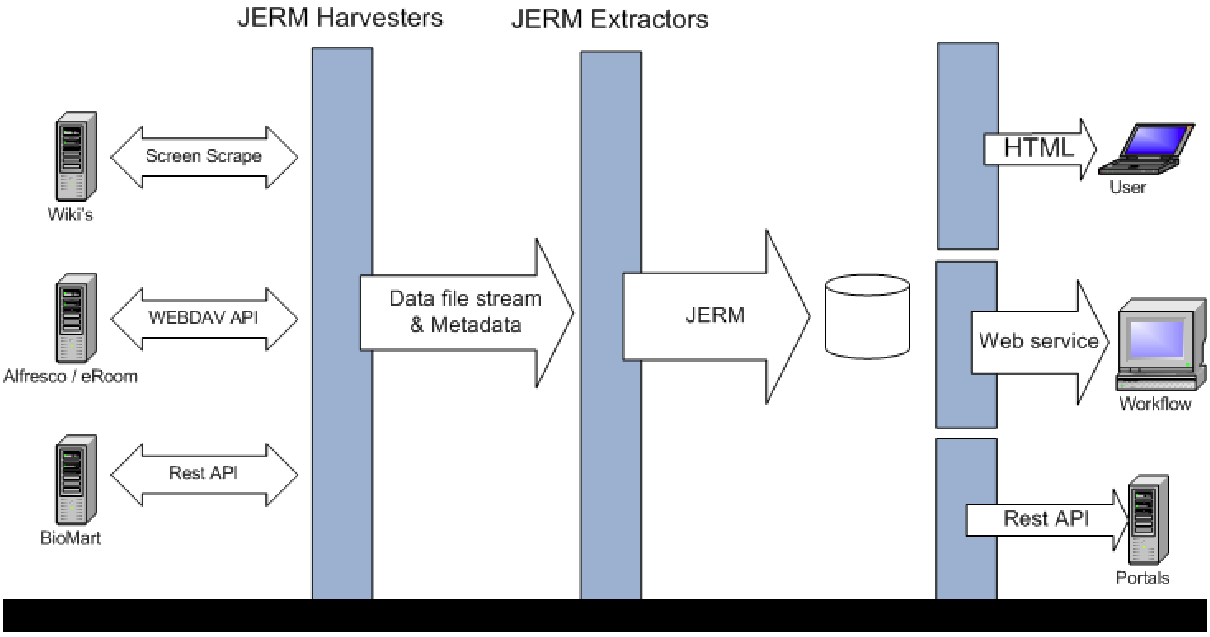
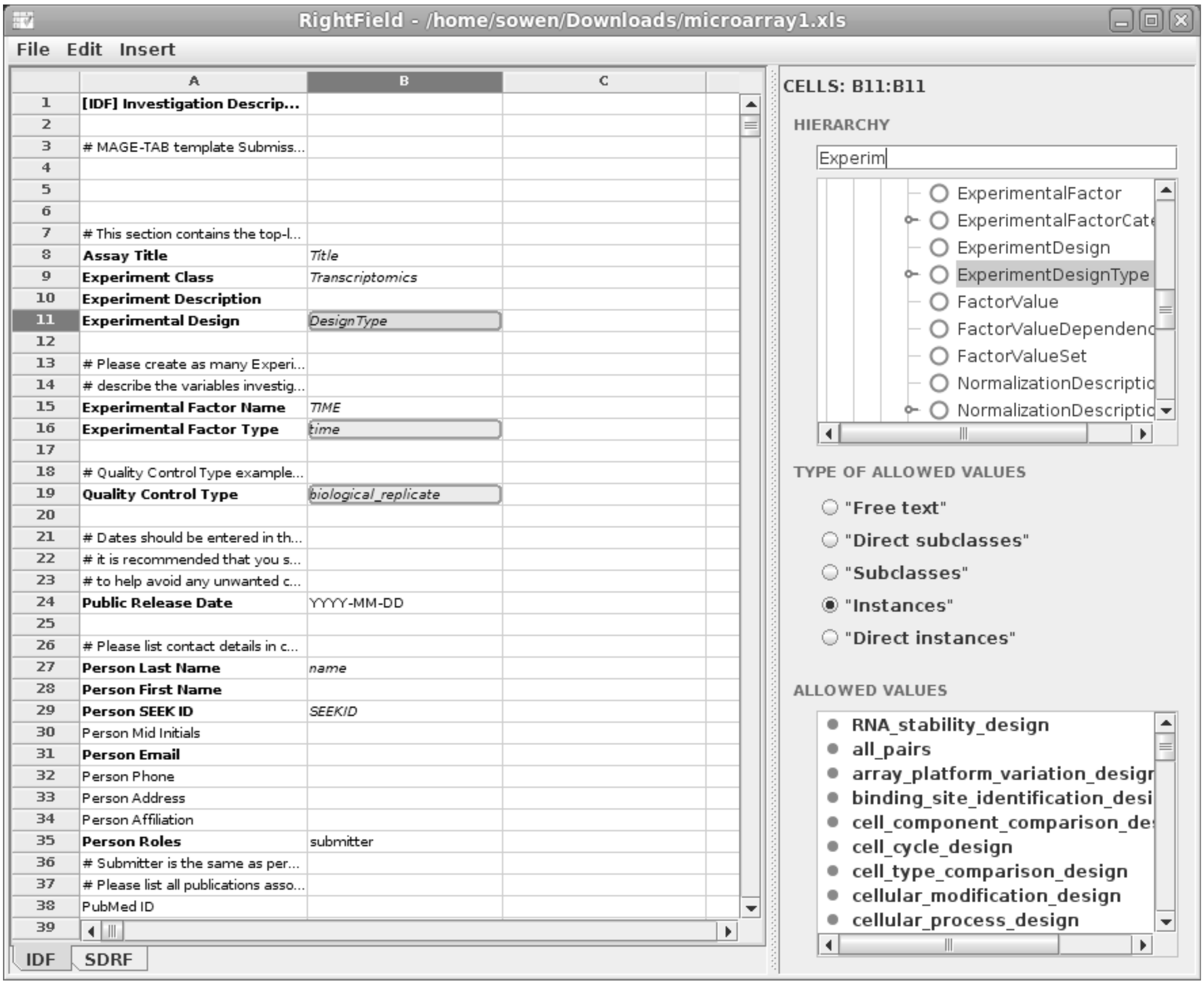


Figure 5: The RightField application showing the embedding of ontology terms into a spreadsheet template from SysMO.



	A	B	C
1	**[IDF] Investigation Descrip...**		
2			
3	# MAGE-TAB template Submiss...		
4			
5			
6			
7	# This section contains the top-l...		
8	**Assay Title**	*Title*	
9	**Experiment Class**	*Transcriptomics*	
10	**Experiment Description**		
11	**Experimental Design**	*DesignType*	
12			
13	# Please create as many Experi...		
14	# describe the variables investig...		
15	**Experimental Factor Name**	*TIME*	
16	**Experimental Factor Type**	*time*	
17			
18	# Quality Control Type example...		
19	**Quality Control Type**	*biological_replicate*	
20			
21	# Dates should be entered in th...		
22	# it is recommended that you s...		
23	# to help avoid any unwanted c...		
24	**Public Release Date**	YYYY-MM-DD	
25			
26	# Please list contact details in c...		
27	**Person Last Name**	*name*	
28	**Person First Name**		
29	**Person SEEK ID**	*SEEKID*	
30	Person Mid Initials		
31	**Person Email**		
32	Person Phone		
33	Person Address		
34	Person Affiliation		
35	**Person Roles**	submitter	
36	# Submitter is the same as per...		
37	# Please list all publications asso...		
38	PubMed ID		
39			

CELLS: B11:B11

HIERARCHY

Experim

- ○ ExperimentalFactor
- ○ ExperimentalFactorCate
- ○ ExperimentDesign
- ○ ExperimentDesignType
- ○ FactorValue
- ○ FactorValueDependenc
- ○ FactorValueSet
- ○ NormalizationDescriptic
- ○ NormalizationDescriptic

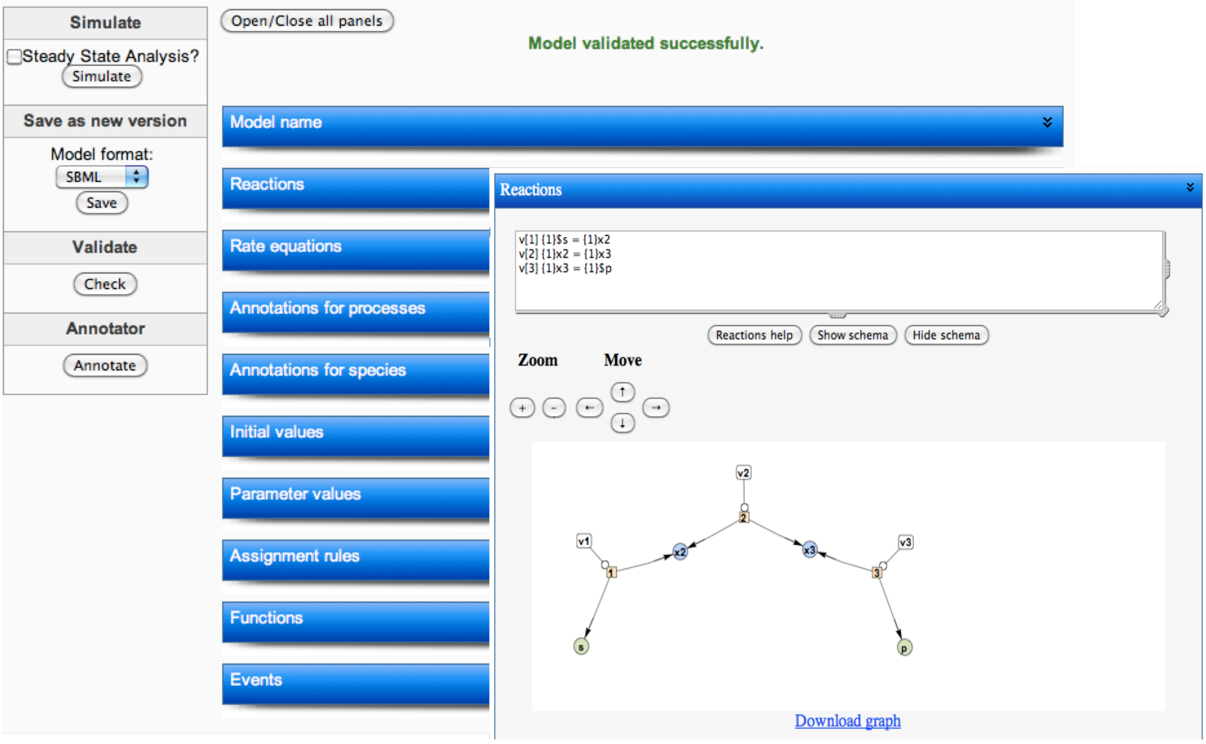
TYPE OF ALLOWED VALUES

○ "Free text"
○ "Direct subclasses"
○ "Subclasses"
● "Instances"
○ "Direct instances"

ALLOWED VALUES

- RNA_stability_design
- all_pairs
- array_platform_variation_desigr
- binding_site_identification_desi
- cell_component_comparison_de
- cell_cycle_design
- cell_type_comparison_design
- cellular_modification_design
- cellular_process_design

Figure 6: The OneStop model constructor. Via a number of text files a user can define a mathematical model, which can subsequently be simulated via the JWS interface. Models can be defined from scratch or uploaded from the JWS Online database or Biomodels database in SBML format. Good error catching and graphical displays of reaction networks (in SBGN format) and rate equations enhance the functionality of the model constructor.

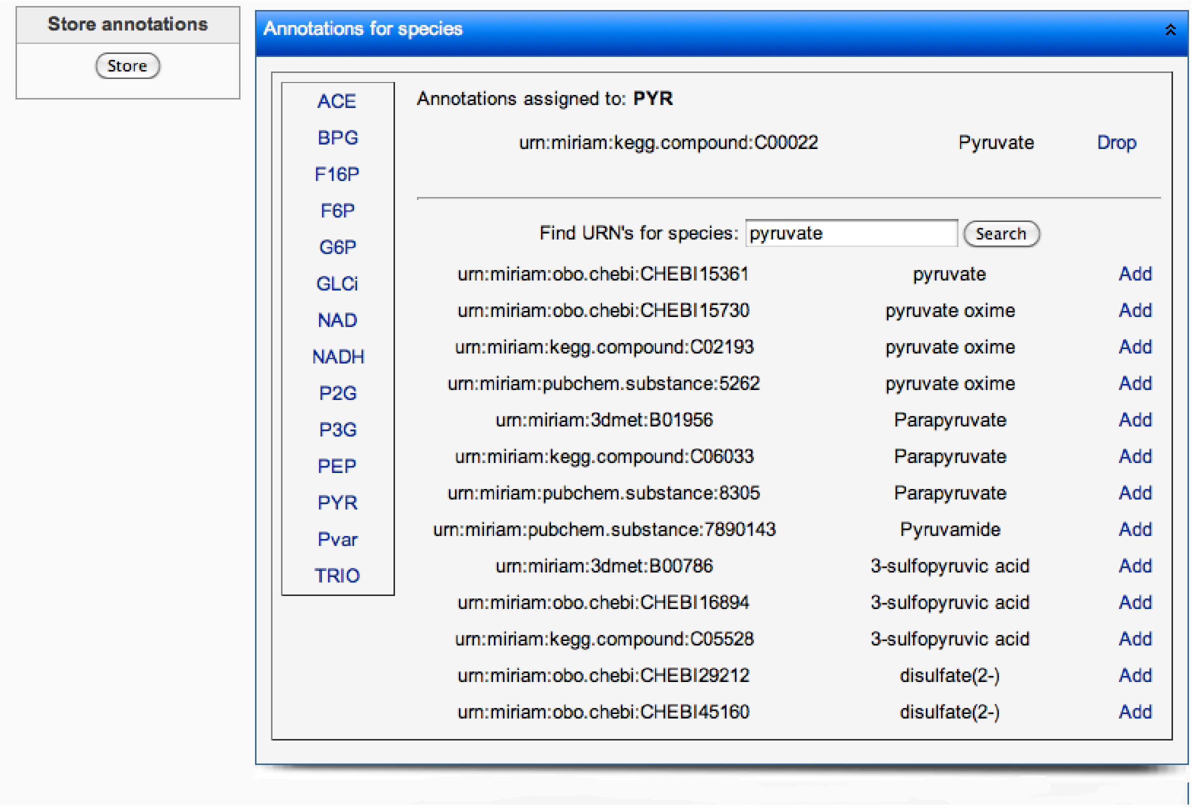


Model validated successfully.

Model format: SBML

```
v[1] {1}$s = {1}x2
v[2] {1}x2 = {1}x3
v[3] {1}x3 = {1}$p
```

Figure 7: The OneStop model annotator. Using the semanticSBML (Krause et al., 2009) webservices, unique identifiers can be given to model variables and model reactions. OneStop makes it possible to annotate the model according to the MIRIAM specification (Le Novere et al., 2005).



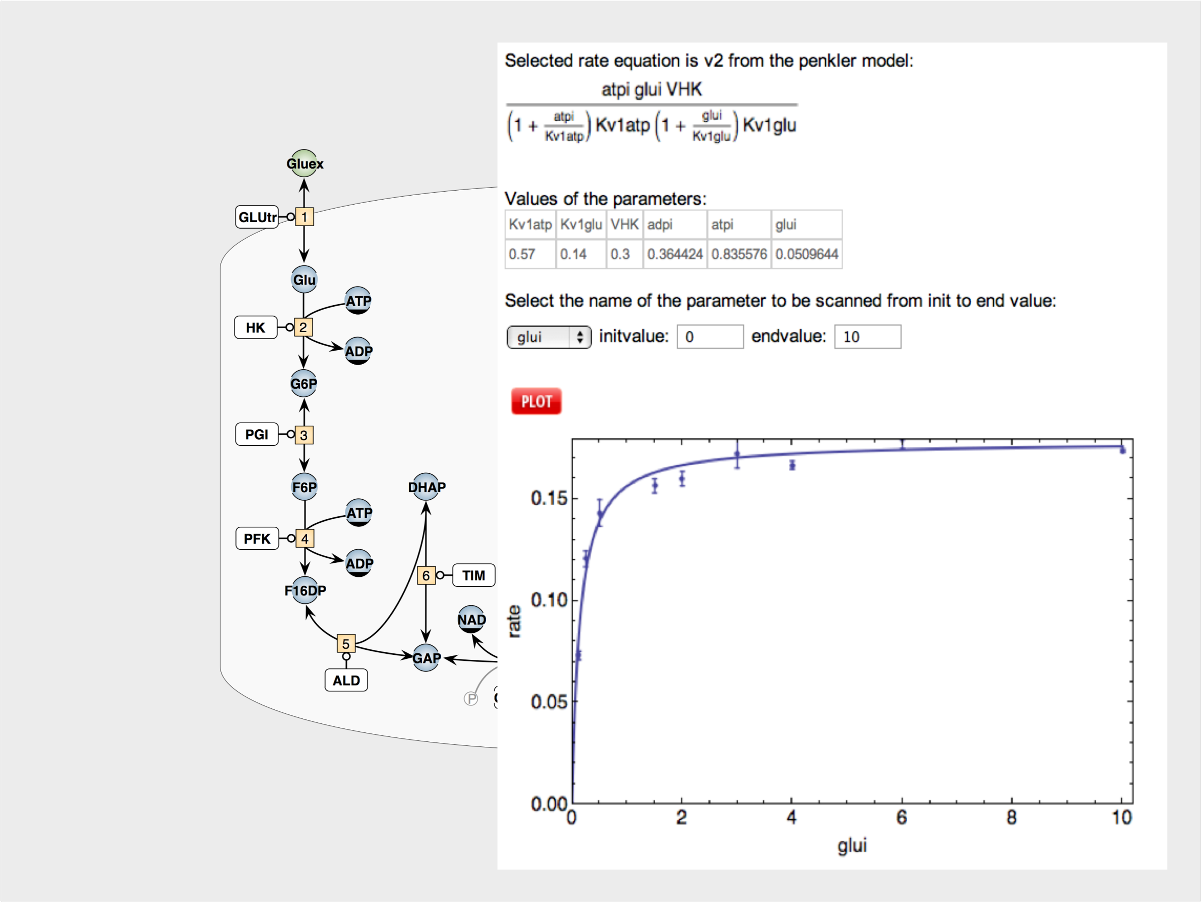
Annotations assigned to: **PYR**

urn:miriam:kegg.compound:C00022 Pyruvate Drop

Find URN's for species: pyruvate

urn:miriam:obo.chebi:CHEBI15361	pyruvate	Add
urn:miriam:obo.chebi:CHEBI15730	pyruvate oxime	Add
urn:miriam:kegg.compound:C02193	pyruvate oxime	Add
urn:miriam:pubchem.substance:5262	pyruvate oxime	Add
urn:miriam:3dmet:B01956	Parapyruvate	Add
urn:miriam:kegg.compound:C06033	Parapyruvate	Add
urn:miriam:pubchem.substance:8305	Parapyruvate	Add
urn:miriam:pubchem.substance:7890143	Pyruvamide	Add
urn:miriam:3dmet:B00786	3-sulfopyruvic acid	Add
urn:miriam:obo.chebi:CHEBI16894	3-sulfopyruvic acid	Add
urn:miriam:kegg.compound:C05528	3-sulfopyruvic acid	Add
urn:miriam:obo.chebi:CHEBI29212	disulfate(2-)	Add
urn:miriam:obo.chebi:CHEBI45160	disulfate(2-)	Add

Figure 8: Example of experimental data linking to individual rate equation. Saturation of the hexokinase reaction with internal glucose is shown as an example for the linking of individual rate equations with experimental data. When the user clicks the reaction process (v2) in the SBGN schema, the rate equation of the reaction is loaded from the model package. The user can select to plot the rate equation as a function of its parameters (here glui for internal glucose was selected). If data for the specific model and rate equation are available in the JWS database these are selected and plotted together with the rate equation.



Selected rate equation is v2 from the penkler model:

$$\frac{\text{atpi glui VHK}}{\left(1+\frac{\text{atpi}}{\text{Kv1atp}}\right)\text{Kv1atp}\left(1+\frac{\text{glui}}{\text{Kv1glu}}\right)\text{Kv1glu}}$$

Values of the parameters:

Kv1atp	Kv1glu	VHK	adpi	atpi	glui
0.57	0.14	0.3	0.364424	0.835576	0.0509644

Select the name of the parameter to be scanned from init to end value:

glui initvalue: 0 endvalue: 10

Tables

	Preferred name	**Synonyms**	**ID**
Gene Name	ALDOA	ALDA GSD12 MGC10942 MGC17716 MGC17767	NCBI-GI: 4557305 NCBI-GeneID: 226 HGNC: 414 HPRD: 00070 Ensembl: ENSG00000149925 KEGG: hsa226
Protein Name	Fructose-bisphosphate aldolase A	Lung cancer antigen NY-LU-1 Muscle-type aldolase	Uniprot: P04075 PIR: S14084
Enzyme classification	Aldolase A, fructose-bisphosphate		EC:4.1.2.13

Table 1: The names and synonyms of a gene and its product in different Life Science databases.

Data	**MIBBI Model**	**Ontologies**	**Standards Body**
Microarray	MIAME:Minimum Information about a Microarray Experiment (Brazma et al, 2001)	MGED (Whetzel et al, 2006)	Functional Genomics Data Society
Proteomics	MIAPE: Minimum Information about a Proteomics Experiment (Taylor et al, 2007)	PSI-MI, PSI-MS, PSI-MOD	Proteomics Standards Initiative
Interaction experiments	MIMIX:Minimum Information about a Molecular Interaction Experiment (Orchard, et al, 2007)	PSI-MI Protein-Protein Interaction	Proteomics Standards Initiative
Systems Biology Models	MIRIAM:Minimal Information Required In the Annotation of biochemical Models (Le Novere et al, 2007)	SBO: Systems Biology Ontology	BioModels.net
Systems Biology Model Simulation	MIASE:Minimum Information About a Simulation Experiment	KISAO:Kinetic Simulation Algorithm Ontology	BioModels.net

Table 2: A selection of minimum information models and their accompanying biological ontologies