Systems biology informatics for the development and use of genome-scale metabolic models

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Engineering and Physical Sciences

Neil Swainston

School of Computer Science

2011

Table of Contents

Abstract **3**

Declaration **4**

Candidate particulars **6**

List of publications **7**

Summary of aims and achievements **9**

1 Introduction 9

1.1 The case for systems biology 9
1.2 A renewed interest in metabolism 11
1.3 Representing metabolism: genome-scale network reconstructions 13
1.4 Mathematical biology: modelling 14
1.5 Aims and achievements 16

2 Genome-scale metabolic networks 19

2.1 Overview of metabolic networks 19
2.2 Computational representation of metabolic networks 21
2.3 Development and applications of a consensus metabolic reconstruction of *Saccharomyces cerevisiae* 22
2.4 From reconstruction to model 25
2.5 Existing approaches in the development of draft metabolic networks 28
2.6 Supporting the development of genome-scale metabolic networks 30
2.7 Conclusions 37

3 Experimental data and kinetic modelling 39

3.1 Overview 39
3.2 Supporting enzyme kinetics assays for systems biology 40
3.3 Quantitative proteomics in kinetic modelling 44
3.4 Automation of the generation of kinetic models from genome-scale metabolic networks 50
3.5 Conclusions 52

4 Conclusion 54

4.1 Future requirements for experimental data management 54
4.2 Future requirements for model development and management 57
4.3 Contributions of this work to data integration in metabolic systems biology 58

5 References 61

Publications **81**

Abstract

Systems biology attempts to understand biological systems through the generation of predictive models that allow the behaviour of the system to be simulated *in silico*.

Metabolic systems biology has in recent years focused upon the reconstruction and constraint-based analysis of genome-scale metabolic networks, which provide computational and mathematical representations of the known metabolic capabilities of a given organism. This thesis initially concerns itself with the development of such metabolic networks, first considering the community-driven development of consensus networks of the metabolic functions of *Saccharomyces cerevisiae*. This is followed by a consideration of automated approaches to network reconstruction that can be applied to facilitate what has, until recently, been an arduous manual process.

The use of such large-scale networks in the generation of dynamic kinetic models is then considered. The development of such models is dependent upon the availability of experimentally determined parameters, from omics approaches such as transcriptomics, proteomics and metabolomics, and from kinetic assays. A discussion of the challenges faced with developing informatics infrastructure to support the acquisition, analysis and dissemination of quantitative proteomics and enzyme kinetics data follows, along with the introduction of novel software approaches to address these issues.

The requirement for integrating experimental data with kinetic models is considered, along with approaches to construct, parameterise and simulate kinetic models from the network reconstructions and experimental data discussed previously.

Finally, future requirements for metabolic systems biology informatics are considered, in the context of experimental data management, modelling infrastructure, and data integration required to bridge the gap between experimental and modelling approaches.

Declaration

Candidate Name: Neil Swainston

Faculty: Engineering and Physical Sciences

Thesis Title: Systems biology informatics for the development and utility of genome-scale metabolic models

The candidate is named as first author for publications 3, 4, 5, 8 and 9. The candidate contributed equally to the first author in publications 1, 5 and 7. The candidate contributed to publication 2.

All contributions have been developed and published while the candidate has been employed at the University of Manchester.

The contributions have not been submitted in support of any other degree or qualification of this or any other university or of any professional or learned body.

I confirm that this is a true statement and that, subject to any comments above, the submission is my own original work.

Neil Swainston, November 2011.

Copyright statement

Statement

Candidate particulars

The candidate holds a BSc (Hons) Chemistry with Industrial Experience (University of Manchester; first-class) and an MSc Computing Science (University of Newcastle-upon-Tyne).

The candidate has over five years research experience in the role of Experimental Officer in the Manchester Centre for Integrative Systems Biology (MCISB), and is currently a full-time member of staff at the University of Manchester in the School of Computer Science and the Faculty of Engineering and Physical Sciences.

Within this role, the candidate has published 18 papers in peer-reviewed journals and conference proceedings articles in the field of systems biology informatics and data management.

List of publications

1. A consensus yeast metabolic network obtained from a community approach to systems biology. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB. *Nat. Biotechnol*. 2008, **26**, 1155-60.
2. Further developments towards a genome-scale metabolic model of yeast. Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD, Mendes P. *BMC Syst Biol*. 2010, **4**, 145.
3. libAnnotationSBML: a library for exploiting SBML annotations. Swainston N, Mendes P. *Bioinformatics.* 2009, **25**, 2292–3.
4. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. Swainston N, Smallbone K, Mendes P, Kell DB, Paton NW. *J Integr Bioinform*. 2011, **8**, 186.
5. SBML Level 3 Package Proposal: Annotation. Waltemath D, Swainston N, Lister A, Bergmann F, Henkel R, Hoops S, Hucka M, Juty N, Keating S, Knuepfer C, Krause F, Laibe C, Liebermeister W, Lloyd C, Misirli G, Schulz M, Taschuk M, Le Novère N. *Nature Precedings*. 2011 http://dx.doi.org/10.1038/npre.2011.5610.1.
6. Enzyme kinetics informatics: from instrument to browser. Swainston N, Golebiewski M, Messiha HL, Malys N, Kania R, Kengne S, Krebs O, Mir S, Sauer-Danzwith H, Smallbone K, Weidemann A, Wittig U, Kell DB, Mendes P, Müller W, Paton NW, Rojas I. *FEBS J*. 2010, **277**, 3769–79.
7. An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ. Siepen JA, Swainston N, Jones AR, Hart SR, Hermjakob H, Jones P, Hubbard SJ. *Proteome Sci.* 2007, **5**, 4.

8. A QconCAT informatics pipeline for the analysis, visualisation and sharing of absolute quantitative proteomics data. Swainston N, Jameson D, Carroll K. *Proteomics.* 2011, **11**, 329–33.

9. Integrative Information Management for Systems Biology. Swainston N, Jameson D, Li P, Spasić I, Mendes P, Paton NW. *In proceedings of the 7th International workshop on Data Integration in the Life Sciences 2010 (DILS'10), Gothenburg, Sweden. Lecture Notes in Computer Science*. 2010, **6254**, 164-178.

Summary of aims and achievements

1 Introduction

1.1 The case for systems biology

With the introduction of whole-genome sequencing, the field of biology entered the post-genomic era, and with this came the promise that the phenotypic behaviour of an organism could be predicted from computational analysis of its genome, with ramifications for disease prediction and treatment, drug discovery and industrial bioengineering. Genome analysis allows the collection of genes that a given organism is capable of expressing to be determined. Coupled with complementary "omics" technologies such as transcriptomics [1], proteomics [2] and metabolomics [3], such genome-scale approaches allow the cellular molecular components of a biological sample to be routinely measured. Genomic sequencing and related omics technologies have therefore greatly increased the understanding of the *potential* of an organism, essentially providing a parts list of genes, mRNA transcripts and proteins that can be expressed, translated and function within the cell.

Much of the biology of the last century followed a qualitative, reductionist approach, based on the theory that the complexity of life could be understood by studying the physiochemical properties of the individual molecules that constitute the cell. While such an approach has provided a wealth of biological knowledge and experimental techniques that continue to be enhanced and exploited today, it is now recognised that it is only through a more quantitative, holistic study of the *dynamic interactions* of these molecules that the function of the organism can be fully elucidated [4]. Such emergent properties cannot be predicted from consideration of the individual molecules alone: the dynamic properties of a cellular system are greater than the sum of its parts. Large-scale omics technologies and the findings of molecular biology remain invaluable data sources upon which more holistic studies can be built. However, neither field provides a quantitative, predictive means of determining system level properties based on either a molecular parts list or a detailed, physiochemical understanding of individual molecules. It is worth considering an initial commentary [5] that accompanies the

publication of the publicly funded human genome sequence [6], which states that "understanding what does give us our complexity remains a challenge for the future". It is this understanding of complexity that systems biology attempts to address.

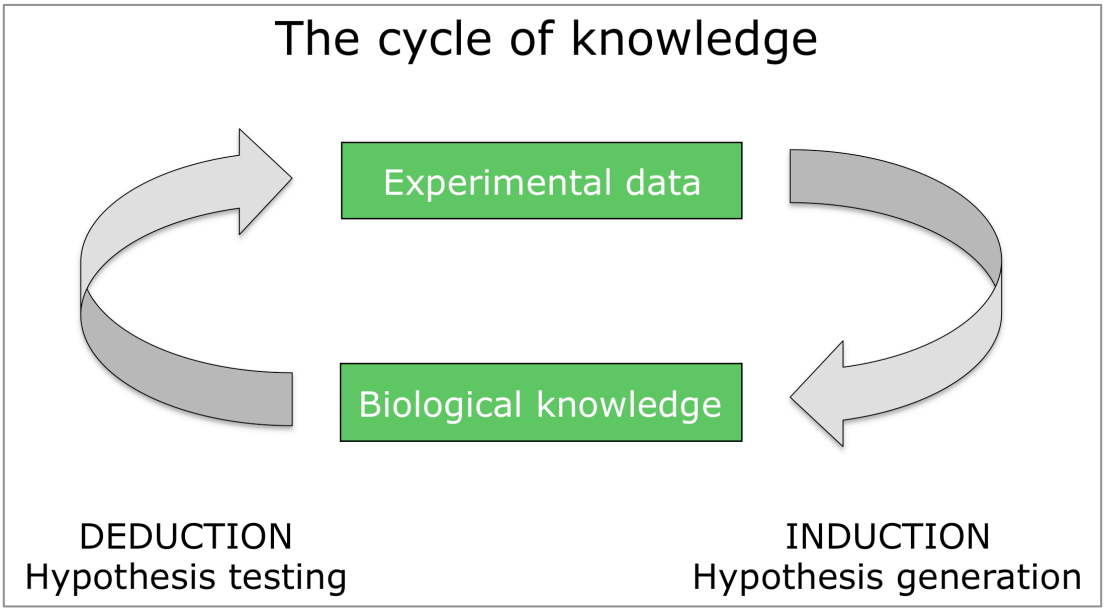

Figure 1: The cycle of knowledge forms an iterative inductive/deductive cycle. The acquisition and analysis of experimental data drives the generation of hypotheses, which can then be tested by generation of subsequent data. Such a paradigm is typical of systems biology studies, in which quantitative modelling of large-scale omics datasets may be used to generate understanding of the biological system. Crucially, computation underpins both experimental data analysis and hypothesis generation through mathematical modelling.

Systems biology [7,8] takes an interdisciplinary approach to studying life, applying existing engineering principles and drawing upon diverse fields in the natural and life sciences in order to understand the inherently non-linear behaviour of biological systems. Systems biology has the ultimate goal of developing quantitative models to allow the behaviour of a biological system to be predicted under a given set of conditions. While the concept of systems behaviour in biology is not new [9], it is only the advent of more powerful computing resources, along with the development of genome-scale omics technologies, that has allowed the field of systems biology to grow as its own discipline over the last ten years.

Due to its utilisation of omics technologies, systems biology can be thought of as a *data-driven* science: data is generated and analysed, and from this hypotheses are generated [10]. This contrasts with the reductionist approach that is hypothesis-driven: a hypothesis is generated which is then either proved or disproved by experiment. In principle, however, systems biology encompasses both data- and hypothesis-driven approaches, ideally forming an inductive and deductive iterative cycle, in which large-scale data sets are generated and analysed (through modelling) to generate hypotheses, which are then tested against subsequently generated data sets (see Figure 1). Such a paradigm is heavily reliant on computation, in the management and analysis of experimental data, the generation of quantitative models, and the integration of data and models. This thesis attempts to address some of the challenges faced in driving this iterative cycle of knowledge.

It is clear that complex organisms such as humans consist of numerous subsystems, including the circulatory and lymphatic systems. While some projects attempt large scale physiological modelling of such systems, or individual organs [11], the majority of approaches currently focus upon modelling the biochemistry of single cells. Cells themselves are characterised by a number of integrated subsystems, including metabolic networks, signalling pathways, transcription networks, and the cell cycle. The boundaries between such subsystems are largely conceptual, illustrated by the discovery of multifunctional proteins that can act as both metabolic enzymes and transcription factors [12], and the knowledge that a highly connected metabolite such as ATP is likely to participate in every cellular subsystem. Ideally, approaches that consider the integration of each, resulting in comprehensive, integrative models that more completely represent cellular behaviour as a whole, will continue to be developed [13]. However, this work will concentrate on the computational approaches required for modelling metabolic networks.

1.2 A renewed interest in metabolism

Metabolism is the integrated network of chemical reactions that take place within the cell. Metabolism is commonly considered as two complementary processes: catabolism and anabolism [14]. Catabolism describes the process of breaking down nutrients from the extracellular environment to extract energy and produce the

organic raw materials necessary for cellular growth. Anabolism is the collection of reactions that use this energy and building blocks to synthesise required cellular components, such as DNA and proteins. Metabolism is common to all life forms, and many of its individual pathways have been well conserved throughout evolutionary history. Studying the metabolic processes of primitive, single-cellular organisms can therefore allow the inference of orthologous functions in higher organisms such as humans.

Although metabolism has long been studied (the word itself is thought to have first been used in the 14th Century [15]), interest in the field waned as interest in genetics grew in the latter half of the last century. Recently, however, the importance of metabolism has again been recognised, due to the central role that it plays in many human diseases and its utility in industrial biotechnology [16].

The presence of altered metabolic states in cancer cells has long been reported, with the Warburg effect – the preference of the tumour cell to excrete lactate rather than exploit its energetic value through oxidative phosphorylation – being well known [17]. Despite the current incomplete understanding of cancer metabolism, many compounds targeting tumour metabolism are in clinical trial [18]. The role of metabolism in other diseases of increasing importance in an overweight and aging society has also been reported [19,20]. Metabolism is generally thought of as being “downstream” from gene expression, protein translation, and enzymatic action, and therefore the metabolome is more indicative of a given phenotype than either the transcriptome or the proteome. It is consequently thought that a deeper understanding of the metabolic state of diseased cells, generated through systems biology approaches, will facilitate and speed up the development of more effective therapies with reduced side effects [21,22].

The study of metabolism is increasing in importance in industrial biotechnology, with metabolic engineering growing in importance. It is clear that microorganisms such as *S. cerevisiae* have long been utilised in the industrial production of naturally occurring products such as alcohols. However, in recent years these techniques have been extended to microbial production of fine chemicals, drugs and biofuels [23]. Such approaches, which are now competing with the more traditional

production technique of synthetic organic chemistry, are essentially efforts in systems biology (or its related discipline, synthetic biology [24]). The system itself is typically a microorganism that has been engineered to maximise production of a required product from selected (preferably inexpensive and sustainable) starting materials. The modified organism may include specific enzymes or entire metabolic pathways from other organisms, or indeed, entirely novel synthetic enzymes that exhibit required metabolic functions. Such microbes are typically modified versions of widely used hosts such as *S. cerevisiae* or *E. coli*. Consequently, quantitative, predictive models of the metabolism (and signalling pathways and transcription factor networks) of these and other host organisms will prove to be an invaluable contribution from the field of systems biology.

1.3 Representing metabolism: genome-scale network reconstructions

The integrated network of chemical reactions that constitute metabolism can be represented as a genome-scale metabolic network reconstruction. Such reconstructions define the network in terms of individual reactions, consisting of metabolic reactants and products, and – where possible – the enzyme, isoenzymes or protein complexes that are known to catalyse the reactions. Additional details to be added include specification of the reaction stoichiometry and directionality. Reaction directionality can be automatically inferred from thermodynamic consideration of the reaction participants [25-27]. In the case of eukaryotic organisms, a definition of the subcellular compartment(s) in which the reaction occurs is required, along with specification of metabolic transport reactions.

The earliest, pre-genome era reconstructions were generated for well-studied microorganisms such as *E. coli*, for which there was sufficient literature describing individual enzymatic reactions that allowed the construction of a coherent network [28]. More recent approaches [29] have utilised collections of metabolic reactions held in curated data sources such as KEGG [30] and MetaCyc [31]. With the prevalence of whole genome sequencing, metabolic reconstructions can now be generated for less well-studied organisms, as homology searching of the sequence itself can identify enzymatic genes from which metabolic functions can be inferred

from better characterised organisms. As such, initial draft metabolic reconstructions can now be automatically generated from genome sequences alone [32].

The range of organisms covered by genome-scale metabolic reconstruction is increasing, partially due to automated methods, such that examples from all three domains of life (archaea, bacteria, and eukaryote) are now represented [33]. The applications of such reconstructions also continue to grow [34], and now cover metabolic engineering, gene annotation, phenotype prediction, network analysis, and bacterial evolution studies. Metabolic reconstructions are therefore a well-established tool in the systematic study of metabolism. In some sense, metabolic networks can never be thought of as being "finished": their development follows an iterative process; with incrementally updated versions being released following newly published experimental data and ever more comprehensive literature describing novel biochemical knowledge [28]. As such, the continued support of their development remains a priority.

1.4 Mathematical biology: modelling

Mathematical modelling is an essential aspect of systems biology. In fact, it may be the use of mathematical modelling that most clearly differentiates systems biology from the molecular biology approaches of the past. The use of modelling in engineering – whether in the design of software, aircraft or buildings – is widespread and, in many cases, essential. The differing mind-set of biologists and engineers has been reported [35], and until recently, mathematical modelling in the study of biological systems remained the concern of a select few researchers [36,37].

A mathematical model provides a quantitative representation of a given system, upon which simulations can be performed to predict an outcome. Typically, the modelling process is an iterative one that involves developing a model that accurately predicts a known set of outcomes upon simulation. Upon production of a predictive model, the model can simulate the effect of manipulations, allowing unknown outcomes to be predicted without the need for additional experimentation. In a biological context, a mathematical model can be used to predict the effect of a gene knockout, or could be analysed to determine which

aspects of the model (reactions or pathways) contribute most to increasing the yield of an industrially relevant product. Modelling in systems biology can take a number of forms, but two approaches – constraint-based modelling and deterministic kinetic modelling – will be considered in this work.

Constraint-based modelling applies linear programming with techniques such as flux balance analysis (FBA) to analyse the flow of metabolites through a metabolic network at steady state [38]. By determining intracellular metabolic fluxes, one can predict the growth rate of an organism, gain an understanding of a given phenotype, or generate hypotheses of how the growth media or the network itself may be altered to exhibit a desired property. FBA can be performed on genome-scale stoichiometric networks, and in principle relies on no additional experimental data in addition to that used to construct the network itself.

While constraint-based modelling has its advantages – not least, its ability to perform analyses on genome-scale models without the requirement of detailed experimental parameters – it is recognised that the behaviour of the cell (and components within the cell) is dynamic. Such dynamics are not captured by constraint-based approaches, which simulate the system at steady state. To predict dynamic behaviour, such as the variance in metabolite concentrations over time, kinetic modelling is required [39]. Due to the increased complexity of kinetic modelling, the systems considered are typically smaller than those in constraint-based modelling, often focusing on a single metabolic pathway [40]. One approach to kinetic modelling is to again represent the system as a collection of individual metabolic reactions, but with each reaction being defined as a non-linear kinetic rate law, with the rate of change of the reaction participant concentrations over time being defined by ordinary differential equations (ODEs). By solving the collection of ODEs, the dynamic behaviour of the metabolites may be simulated.

Another key difference between constraint-based and kinetic modelling is the latter's dependency upon experimentally derived kinetic parameters. While many such kinetic parameters have already been measured and are publicly available in resources such as BRENDA [41] and SABIO-RK [42], the range of data available is currently insufficient for genome-scale kinetic modelling to be realised. Furthermore, the field of enzymology has generated much of the existing data.

While enzymology is a mature field of study, it does not necessarily generate data that is suitable for systems biology studies. Enzymology has differing goals to systems biology, with many studies attempting to optimise enzymatic activity, rather than to determine the properties of enzymes in physiologically relevant conditions. As such, consistent sets of kinetic parameters measured at conditions of the system under study – or even from an appropriate organism – are typically unavailable.

Additionally, the kinetic laws represented by ODEs have a dependency upon enzyme concentration. Although some enzyme concentrations have been measured and are publicly available [43-45], the coverage of such data sets is even smaller than that of kinetic parameters. Consequently, experimental determination of both kinetic parameters and enzyme concentrations remain necessary steps in the development of accurate kinetic models.

1.5 Aims and achievements

The aim of this thesis is to automate the construction of genome-scale metabolic reconstructions, and the production of kinetic metabolic models from quantitative proteomics and enzyme kinetic assay experimental data.

The development of such kinetic models is dependent upon computational support in both the areas of experimental data analysis and management, and in development of the models themselves. As the acquisition of experimental data is expensive in terms of both time and money, the approach followed is to target individual metabolic pathways that are found to carry the greatest metabolic flux under standard conditions.

The determination of metabolic fluxes requires a genome-scale model upon which constraint-based analyses can be performed. From this, individual metabolic pathways can be targeted for experimental work, allowing the determination of enzyme and metabolite concentrations and kinetic constants necessary for kinetic model parameterisation. The final step concerns the integration of experimental data with unparameterised models.

The computational aspects of this pipeline – from reconstruction generation through experimental data analysis and management to automated model parameterisation – will be achieved by fulfilling the following objectives:

- To characterise the steps required in manually constructing metabolic reconstructions through community participation.
- To develop techniques that automate many of these steps, supporting the computational construction of genome-scale metabolic reconstructions from existing data resources.
- To develop techniques for producing kinetic constants from kinetic assay data, such that this data can be utilised in the automated parameterisation of kinetic metabolic models.
- To develop techniques for producing enzyme concentrations from quantitative proteomics data, using the QconCAT approach [46], such that this data can be utilised in the automated parameterisation of kinetic metabolic models.

In fulfilling these objectives, this thesis makes the following contributions:

- Genome-scale metabolic reconstructions of *Saccharomyces cerevisiae* (Publication 1 [73]; Publication 2 [85]), produced through manual construction with community participation. This work represents the first instance of a community-developed reconstruction, and the first instance of a genome-scale reconstruction that adheres to community-developed data standards.
- Informatics support for the automated generation of genome-scale metabolic reconstructions, based on experience gained from the manual approach to metabolic network reconstruction (Publication 3 [98]; Publication 4 [103]). This work is comprised of a modular system that can be used in automated pipelines for the generation of draft reconstructions, and also in curation and refinement of existing reconstructions.
- An integrated approach to analysis, storage and dissemination of kinetic assay experimental data (Publication 6 [114]). This work introduces the first reported method for storage of raw experimental data from kinetic assay experiments, and the first reported method for automated submission of kinetic parameters

to publicly available data resources. This approach was followed in such a way to allow its integration into an automated pipeline for generation of kinetic models (Publication 9 [145]).

- An integrated approach to analysis, storage and dissemination of quantitative proteomics experimental data (Publication 7 [131]; Publication 8 [143]). This work entails the first generation of quantitative proteomics data that adheres to community data standards. This approach was followed in such a way to allow its integration into an automated pipeline for generation of kinetic models (Publication 9 [145]).

2 Genome-scale metabolic networks

This section provides an overview of metabolic networks, expanding on the introduction given in Section 1.3. Computational representations of genome-scale metabolic reconstructions are discussed followed by a specific case study on their community-driven development and subsequent applications. Finally, existing computational approaches for the development of genome-scale metabolic networks are discussed followed by novel techniques that were developed in support of this thesis.

2.1 Overview of metabolic networks

Genome-scale metabolic networks can be broadly considered to take two forms: reconstructions and models [47]. Metabolic reconstructions follow a "bottom-up" approach to capturing a biochemical representation of metabolic functions, and can be as simple as a collection of metabolites and stoichiometric metabolic reactions. Such reconstructions can be thought of as compendiums of biochemical knowledge, and as such may contain information for which there is limited evidence, but may be retained in the reconstruction with the intention of subsequent evaluation. As such, recent reconstructions have focused upon applying confidence levels to represented biochemical information, and – where possible – annotating such information with literature references to support its inclusion. Reconstructions provide a representation of biochemical knowledge similar to those provided by the existing data resources of KEGG [30] and MetaCyc [31]. As such, these resources are often used as sources for initial drafts of metabolic reconstructions. In addition, metabolic reconstructions can be generated from an organism's genome sequence, an inference process involving homology searching of the sequence to find genes encoding for enzymes whose metabolic functions can be exhibited by the organism.

Reconstructions of eukaryotic organisms may be compartmentalised, such that individual reactions are specified to take place in one or more intracellular compartments [48]. Generating a compartmentalised model also necessitates the specification of transport reactions, allowing metabolites to either diffuse or be actively transported across cellular membranes. In the case of active transport,

transport proteins (or complexes) should be specified, again allowing investigation of gene deletions involving these transporters.

Models can be generated from reconstructions and analysed using constraint-based approaches. Additional requirements for models, over and above those of reconstructions, include:

- mass and charge balancing of reactions. Balancing reactions ensures stoichiometric consistency within the network, preventing "leaks" of metabolites, that is, reactions or groups of incorrectly specified reactions that either spontaneously produce or consume mass, violating mass conservation [49].
- definition of reaction directionality. Defining reaction directionality increases the predictive accuracy of the model by preventing reactions from running in a thermodynamically infeasible direction. The determination of reaction directionality can be performed through estimation of the standard Gibbs free energy of formation ($\Delta_f G'^{\circ}$) of each of the metabolic participants, from which the Gibbs free energy of the reaction ($\Delta_r G'^{\circ}$) can be calculated [25-27].
- specification of gene-protein-reaction relationships. Each reaction can optionally be associated with one or more genes (and proteins), or protein complexes, which catalyse its activity. Adding such gene-protein-reaction relationships in a metabolic model allows for the prediction of behaviour of the organism under gene deletion studies [50].
- definition of system boundaries. Even for models that do not consider intracellular compartmentalisation, extracellular and intracellular compartments are typically defined. This allows the specification of a growth medium in terms of a collection of extracellular metabolites that may be taken up by the cell, and the definition of "sink" reactions involving the excretion of intracellular metabolites from the cell. The extracellular space therefore acts as the boundary of the system.
- specification of an objective function. Linear-programming approaches such as flux balance analysis (FBA) generate a number of feasible metabolic flux

patterns that the system may exhibit. An objective function – commonly, the definition of a biomass function that simulates cell growth – is specified and optimised in the linear-programming method.

- filling of gaps. Limiting reconstructions to include only reactions for which there is published literature evidence can give rise to gaps in the network, in which collections of reactions are disconnected from a central core. Reconnecting these reactions to the network relies on gap filling, which may be driven by inferring the presence of reactions from those reported in other organisms, or by adding putative, "modelling" reactions that allow the model to provide improved predictions of experimental data results.

The boundaries between reconstructions and models are not well defined, and as such, a reconstruction may also include gene-protein-reaction relationships, for example. However, for a reconstruction to be utilised in modelling approaches, it must fulfil the above requirements.

Additionally, there is currently no specification of the size required for a network to be considered "genome-scale". The development of metabolic networks is an iterative process, and as such, their size increases over time, driven by advances in experimental work and the publication of ever more comprehensive literature. For example, the first published reconstruction of *E. coli* metabolism contained only 14 metabolic reactions [51], a number that has increased to 1387 in the most recent version [52]. There exists a trade-off between scale and accuracy in metabolic networks. Therefore, networks may contain a large number of reactions for which there is little confidence, while more focused networks may be smaller in scale but only contain reactions for which there is literature or experimental evidence. There is also little agreement regarding quality of a metabolic network, although commonly used metrics include a measure of network connectivity, and in the case of models of microorganisms, the ability to predict the results of gene-deletion studies.

2.2 Computational representation of metabolic networks

As systems biology has matured as a field over the last ten years, a number of standard data formats have been developed for allowing the interchange of

models. These include CellML [53], BioPAX [54] and the Systems Biology Markup Language (SBML) [55], and each of these is encoded in eXtensible Markup Language (XML). SBML is the most widely utilised of these formats, and is currently supported by over 200 software tools, covering a range of modelling applications. SBML allows the definition of systems biology models in terms of chemical species, reactions between species, and cellular compartments. As such, it is an appropriate format for the representation of genome-scale metabolic networks. Furthermore, SBML is also applicable to kinetic models, providing the capability for representing mathematical rate laws that describe the dynamic behaviour of individual reactions within the system. A further advantage of SBML is its support by dedicated constraint-based modelling software packages such as the COBRA Toolbox [56], the BioMet Toolbox [57] and FASIMU [58]. Consequently, this thesis will focus on the development and use of metabolic models encoded in SBML.

2.3 Development and applications of a consensus metabolic reconstruction of *Saccharomyces cerevisiae*

Since the development of the first genome-scale metabolic reconstruction of yeast metabolism, iFF708, in 2003 [59], a number of subsequent competing versions have been derived from the original reconstruction. This led to the requirement for the community to develop a consensus reconstruction from existing sources that would reflect the current community knowledge of yeast metabolism.

The project entailed taking two separately developed reconstructions, iMM904 [48] and iLL672 [60], and merging them to form a community-agreed consensus to be utilised and developed further. The development of a consensus was facilitated by inviting interested parties to a “jamboree” [47], a focused meeting in which experts in yeast metabolism, systems biology, metabolomics and information management worked together on individual tasks required in merging the two reconstructions. The main tasks of this meeting were to i) define the scope of the consensus model: what it would contain and how it would be represented; ii) identify and annotate chemical species (metabolites, genes and proteins) within the consensus; and iii) resolve discrepancies between the two original reconstructions.

It was decided that the consensus reconstruction would consist of the following components:

- metabolic reactions;
- gene-protein-reaction relationships specifying the (iso)enzymes, protein complexes and encoding genes that are known to catalyse each metabolic reaction;
- compartmentalisation (both extracellular and intracellular);
- transport reactions, defining how metabolites are transported across the plasma and intracellular membranes; and
- gene-protein-transport reaction relationships specifying the transport proteins and encoding genes that allow for metabolic transport.

Furthermore, the reconstruction was to be generated in SBML, and would unambiguously identify each component within it. The necessity for doing so was immediately apparent given the task in hand: to merge existing models in which individual components were inconsistently named.

The problem of inconsistent naming of chemical elements has been noted as an impediment to performing comparison of models [61]. Rather than relying upon an *ad hoc* definition of naming standards to be used for identifications in the consensus, an existing standard, the Minimum Information Requested In the Annotation of Models (MIRIAM) [62], was used. The MIRIAM standard provides a format for assigning semantic annotations to elements within an SBML model, an example of which is shown in Figure 2.

Identification of chemical species was performed initially by automated and then by manual approaches. In both cases, species names and, in the case of metabolites, chemical formulae were used to query chemical databases such as ChEBI [63] and KEGG. This was augmented with manual searches of scientific literature. In subsequent work on the development of a human consensus reconstruction, identified metabolites that were not present in chemical databases were submitted to ChEBI to be added to the resource [64]. As such, an unexpected contribution of the reconstruction process became the enhancement of third party public resources.

```
<species metaid="meta_glc" id="glc" name="D-Glucose">
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about="#meta_glc">
    <bqbiol:is>
     <rdf:Bag>
      <rdf:li
rdf:resource="urn:miriam:obo.chebi:CHEBI%3A17634"/>
     </rdf:Bag>
    </bqbiol:is>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 2: MIRIAM annotated SBML species representing glucose. The SBML contains a unique id for the species itself, a human readable name, and a block of Resource Description Format (RDF, http://www.w3.org/RDF/) XML, which uses the MIRIAM qualifier `bqbiol:is` and a Uniform Resource Identifier (URI, http://www.w3.org/Addressing/) representing an entry in the ChEBI database to explicitly state that the species is represented by the ChEBI entry `CHEBI:17634`.

The specification of intracellular compartmentalisation, transport reactions and reconciliation of discrepancies between reactions in the two original reconstructions was performed manually through literature searching. In these cases, evidence for a given decision was provided by annotating the reconstruction with appropriate PubMed [65] entries, providing a literature reference to validate the existence of an entry in the consensus.

This work led to the first community-developed consensus metabolic reconstruction of an organism, and the first reconstruction that followed existing standards to unambiguously identify model components through semantic annotations. Cellular compartments were identified with Gene Ontology (GO) terms [66]. Metabolic species were annotated with ChEBI, KEGG, PubChem [67] or HMDB [68] identifiers, and enzymes with the Saccharomyces Database (SGD) identifiers [69] and UniProt terms [70], representing genes and proteins respectively. Reactions were annotated with enzyme classification (EC) terms [71] and PubMed identifiers where possible.

System Biology Ontology (SBO) [72] terms were used to make explicit the distinctions between metabolic and enzymatic species, and between metabolic and transport reactions.

This initial consensus reconstruction is known as "Yeast 1.0". The publication in *Nature Biotechnology* [73] (Publication 1, ISI Impact Factor 31.1) has been highly cited since publication, with 154 citations as of November 2011[i]. Such was the success of the jamboree approach in the development of a consensus reconstruction of yeast metabolism, the approach was also subsequently applied to *Salmonella typhimurium* [74] and human (on-going). The utility of following such a community approach in reconciling existing reconstructions was described in an editorial discussing two competing models of human liver metabolism [61].

Citing articles cover a range of applications, including interpretation of experimental data [45,75], the community-driven curation and expansion of reconstructions [76], and integration with expanded cellular models that consider regulation and protein-protein interactions [77,78], developed through tools such as ONDEX. Such integration of signalling pathways and transcription factor networks with metabolic reconstructions is likely to become increasingly prevalent in future. The first steps in this will involve the reconstruction of genome-scale signalling networks. It is thought that this endeavour can be aided by the experience of developing community driven metabolic reconstructions [79], facilitating the task of generating integrated models of all cellular processes that more completely represent the behaviour of the organism under study.

2.4 From reconstruction to model

A limitation of the initial consensus network (Yeast 1.0) was that, as a reconstruction, it was unable to be used in constraint-based modelling techniques such as FBA. The fundamental limitations of the initial reconstruction in the context of being amenable to FBA were the presence of gaps in the network and the lack of a defined biomass reaction. Furthermore, in terms of the scope of the network, it was seen that metabolite transport and lipid metabolism were both under

[i]Citations are taken from Google Scholar (http://scholar.google.co.uk/)

represented. To rectify this, a number of iterations of the yeast consensus reconstruction were generated, with the fourth iteration being performed by following a jamboree based approach. This iteration involved the amalgamation of a new reconstruction, iIN800 [80], which more comprehensively covered lipid metabolism, and also included a more targeted search for missing enzymes that are present in lipid pathways, utilising LIPID MAPS [81], homology searches of enzymes in KEGG lipid pathways, and manual examination of SGD and Ensembl [82].

Connectivity analysis was performed on the network to find clusters of unreachable metabolites, which were either disconnected from the extracellular medium, or incapable of carrying flux due to the presence of dead-end metabolites (see Figure 3). Such clusters were prioritised in order to focus manual curation efforts on those metabolites whose reconnection would have the greatest impact on the overall network connectivity. This approach was an example of the iterative "cycle of knowledge" advocated by Kell *et al.* [10], in which analysis of a model can be used to either suggest subsequent experimental work, or in this case, drive literature searches to target areas of biological knowledge that are as yet under-represented in existing databases.

A biomass function was added to the reconstruction, which was taken from iIN800, due to its consideration of lipid metabolism. The model was then analysed with FBA for its ability to predict the results of experimentally determined gene knockout experiments [83,84], and was found to correctly predict ability or inability to grow for 80% of gene knockouts – a figure comparable to that of the existing reconstruction iMM904 and greater than that of iIN800.

Again, the model ("Yeast 4.0") was formatted according to the SBML and MIRIAM standards, and the work published in *BMC Systems Biology* [85] (Publication 2, ISI IF 3.6) where it is defined as "Highly Accessed" and has been referenced by 12 publications, including that of Zomorrodi *et al*. [86].

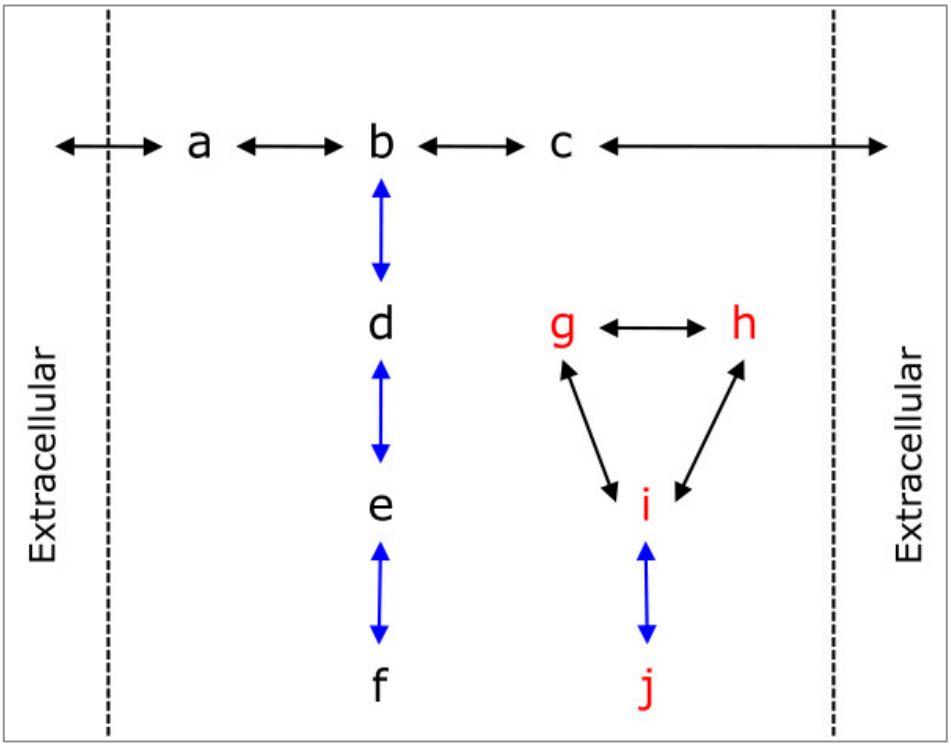

Figure 3: Visualisation of connectivity analysis. Unreachable metabolites (in red) are disconnected from the extracellular medium. "Blocked" reactions (in blue) are incapable of carrying flux as they lead to dead-end metabolites (such as the metabolites f and j). Gap filling is required to reconcile both issues. In the case of the set of unreachable metabolites, each of these would be made reachable if one member of the set were linked to the core network (i.e. by adding any reaction between a black and a red metabolite). Reactions can be unblocked by adding a reaction between a dead-end and a connected metabolite (e.g. a reaction between the metabolites f and c), or by adding a transport reaction that allows the dead-end metabolite to be excreted to the extracellular space.

The goals of Zomorrodi *et al.* were similar to that of Yeast 4.0: to generate a model of yeast metabolism that was capable of better predicting the results of gene knockout experiments. The work of Zomorrodi *et al.* was based on the reconstruction iMM904, and was motivated by the fact that the model's accuracy was significantly worse than existing prokaryotic reconstructions such as that of *E. coli* [52]. The approach applied the existing automated methods GapFill [87] and GrowMatch [88] to resolve false predictions of outcomes of both single and multiple gene deletions from the iMM904 model. The resolving of such false

predictions can include addition of missing metabolic and transport reactions, addition or removal of isoenzymes from individual reactions, relaxation of reaction irreversibility constraints, addition of previously missing metabolites to the *in silico* growth medium, and modification of the biomass objective function. In doing so, the authors suggested 120 modifications to the iMM904 model for which there was literature evidence. While some of these were independently discovered during the development of Yeast 4.0, 86% of them were entirely novel. This highlights both the utility of experimental datasets in the testing of metabolic models, and also the need for their continual refinement.

2.5 Existing approaches in the development of draft metabolic networks

While Section 2.3 discussed the generation of a consensus reconstruction from existing versions, the process of generating reconstructions from scratch has traditionally been an arduous one, with time-scales in the region of six months to two years quoted for their successful production [89].

The first step in the reconstruction process for a given organism is the production of a draft metabolic network, a process that is generally based upon one of two methods. For well-characterised organisms, existing publicly available data resources such as KEGG or MetaCyc may be used. In the case of less well-studied organisms, a putative metabolic network can be inferred from the genome sequence itself, using tools such as SHARKhunt [90] and PathwayTools [91].

Considering the first approach, both KEGG and MetaCyc contain curated representations of metabolism in a range of organisms. While MetaCyc allows for the export of such data in SBML, KEGG does not. However, at least two tools exist to allow reformatting of KEGG data into SBML.

KEGG2SBML (http://sbml.org/Software/KEGG2SBML) provides a simple reformatting of KEGG flat files into SBML. This can be performed for individual pathways, generating a collection of smaller models that can be merged to generate a genome-scale representation of the metabolic capabilities of the organism. However, KEGG2SBML has the disadvantages of being no longer supported, and is incompatible with the current KEGG flat file format. Furthermore, KEGG2SBML periodically fails to include individual metabolic reactions (and indeed entire

metabolic pathways) and as such its use results in metabolic reconstructions that contain significant gaps. KEGGconverter [92] has similar goals to that of KEGG2SBML, in that it claims to convert individual KEGG pathways into SBML. Unfortunately, it has the major disadvantage of ignoring cofactors in its definitions of metabolic reactions, thus generating unbalanced reconstructions that require a great deal of curation to be converted to models that can be used in constraint-based modelling.

Some of these limitations have been overcome by MetNetMaker [93], a graphical tool for developing genome-scale models from reactions specified in KEGG. This tool contains a number of useful features, such as the ability to utilise existing reactions and define novel reactions that are not present in KEGG, the generation of models that are compatible with the COBRA Toolbox, and the provision of a graphical editor. However, it suffers from the limitation of failing to associate its metabolic reactions with enzymes. This prevents any generated models from being used in gene knockout simulations, and prevents the use of resulting networks in analysis of transcriptomic and proteomic data.

MetaCyc allows for exporting of curated metabolic pathways as a single, merged SBML file. However, its export facility also fails to include gene-protein-reaction relationships. While such mappings are available in other downloadable flat files, their incorporation into the metabolic model requires the writing of dedicated parsers.

Another issue with KEGG is the lack of information on intracellular compartmentalisation, and therefore on transport reactions and proteins.

Furthermore, none of the above tools generate semantically annotated models, or use consistent naming of model components, and as such, generating a consensus reconstruction from networks generated from different resources requires each to be reconciled and merged manually. This issue was discussed by Poolman *et al.* [94], who discussed the challenges to be faced in the reconstruction of networks from public databases. Other issues discussed include the presence, in public databases, of unbalanced reactions, duplicate reactions, synonymous metabolites (such as 'trehalose 6-phosphate' and 'α,α-trehalose 6-phosphate') and generic

metabolites (such as 'aldehyde'). As such, any reconstruction generated solely from publicly available databases is likely to contain a number of inconsistencies and fall short of the standard achieved by manual curation.

Perhaps the most sophisticated tool that supports genome-scale network reconstruction is the Model SEED [32], a web application that automates the model building and analysis process from an annotated genome sequence. The Model SEED has been used to generate genome-scale models for 130 bacterial organisms, 22 of which have been optimised and validated against gene knockout datasets. Limitations of the current implementation are that the system is limited to prokaryotic organisms, that generated models do not adhere to the MIRIAM standard, and are therefore not immediately amenable to automated methods of comparison with existing reconstructions, and that there is no support for a subsequent iterative, manual curation process. However, it is understood that support for manual curation is currently in development.

2.6 Supporting the development of genome-scale metabolic networks

The generation of genome-scale metabolic reconstructions for both yeast and human led to the development of a number of software tools to aid their construction, which were ultimately packaged and published.

2.6.1 libAnnotationSBML: an API to support annotated SBML models

The development of software infrastructure to support models encoded in SBML has been greatly facilitated by the development of the programming library libSBML [95]. libSBML provides the facility for parsing and writing SBML models, supporting such tasks as adding and removing SBML elements such as species and reactions, and can also be applied to SBML annotations. Taking the example from Figure 2, libSBML provides the facility to parse the annotation of species elements to extract both the predicate `bqbiol:is` and the URI `urn:miriam:obo.chebi:CHEBI%3A17634`. Furthermore, libSBML allows for the writing of such annotations. However, libSBML provides no further functionality beyond the getting and setting of such terms. There is therefore a requirement to support i) the selection of predicates and appropriate URIs with which to annotate SBML elements; and ii) the interpretation of these terms upon parsing of annotated

models. It is these requirements that libAnnotationSBML attempts to address. libAnnotationSBML provides an API that sits on top of libSBML, and interfaces with the MIRIAM Resources [96] and other web services to aid the reading and writing of semantic annotations in SBML models.

The library can be used in the writing of annotations, exploiting web service search facilities to ChEBI, KEGG, GO and UniProt, allowing elements in unannotated models to be searched against these resources to suggest predicates and URIs that can be used in their annotation.

Once annotated, SBML documents can be parsed, allowing URIs, referencing entries in external data resources, to be extracted from annotated SBML elements and interpreted. libAnnotationSBML separates the URI into data type (a MIRIAM URN, representing the data type to which the URI refers) and id (a data resources specific identifier representing a specific entry in the resource). MIRIAM Resources provides resource descriptions of each data type, both via its web browser (see Figure 4) and web service interface. This web service interface is then queried with the MIRIAM URN, to both determine the data type to which it refers and to check its validity. In addition, MIRIAM Resources provides a number of access URLs, which libAnnotationSBML extracts to provide a mapping from a URI to an accessible URL linking to the specific entry in the data resource (see Figure 4). libAnnotationSBML also interfaces with web services describing the specific entry in external data resources. For example, ChEBI web services provide programmatic access to all of the data shown in Figure 5. Therefore, once libAnnotationSBML has established the identity of a MIRIAM URN pointing to a ChEBI entry, mapping to the appropriate web service calls allows terms such as name, synonyms, chemical formula, charge and InChI string to be automatically extracted, providing a link between the MIRIAM URN and the metadata that describes the term. Such mappings are also available to data resources such as KEGG, PubChem, HMDB, SBO, Gene Ontology (GO), NCBI Taxonomy [97], SGD and UniProt. Cross references between these resources are established where possible, allowing a given ChEBI term to be mapped to its equivalent in KEGG, from which KEGG-specific information, such as reactions in which the metabolite is known to participate, can be extracted.

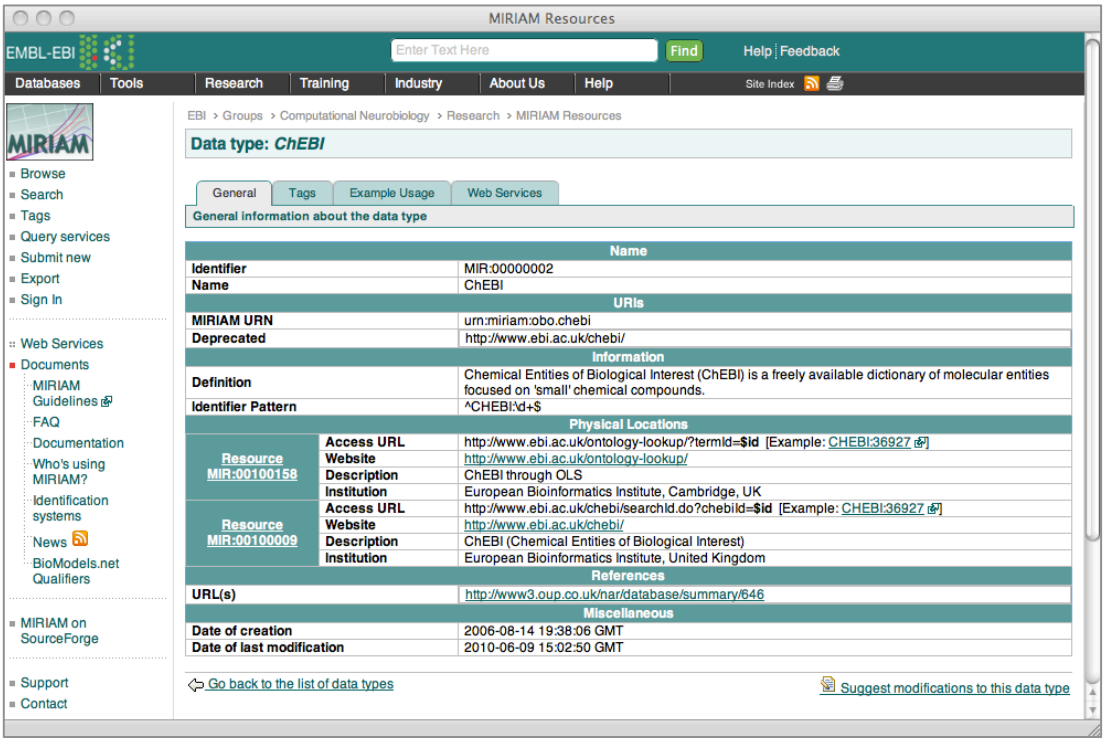

Figure 4: Web browser interface to the MIRIAM Resources entry describing the ChEBI data type.

libAnnotationSBML therefore acts as a link between the semantic annotations captured in an SBML model and the data resources that describe these semantic terms. The library is described in *libAnnotationSBML: an API to support annotated SBML models*, which was published in *Bioinformatics* [98] (Publication 3, ISI IF 4.9), with 11 citations.

Related work focuses on the writing, rather than interpretation, of semantic annotations in SBML models. Such work includes that of Lister *et al.* [99], who have developed a web-based tool, SAINT, to guide the annotation of SBML models. As the tool is focused towards protein-protein interaction pathways, SAINT queries resources relevant to such networks, such as UniProt, STRING [100] and Pathway Commons, to retrieve semantic terms that may be selected for model annotation. In a similar area is the work of Krause *et al.*, semanticSBML [101], which provides the facility for applying MIRIAM annotations to SBML models, but also includes tools to aid their merging.

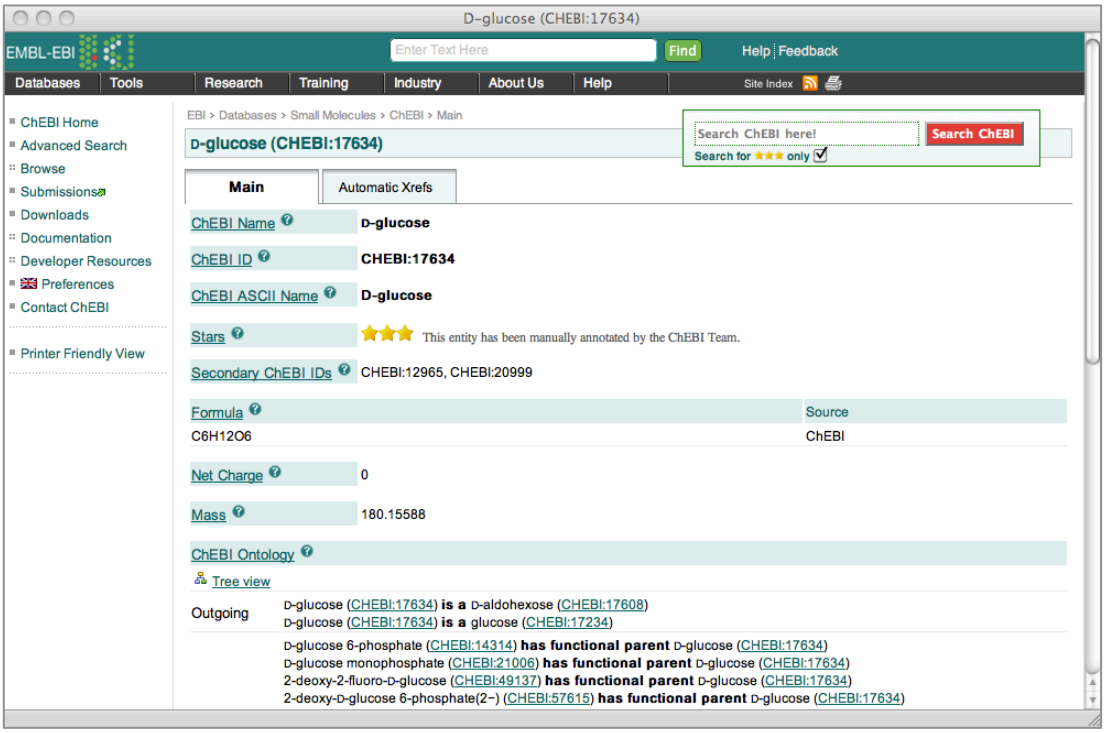

Figure 5: Web browser interface to entry `CHEBI:17634` in the ChEBI database, describing D-glucose.

A primary goal of model annotation is to facilitate their sharing and reuse. Li *et al*. discuss the importance of annotation, and the difficulties associated with providing it, in their paper describing the BioModels Database [102], a publicly available repository for published systems biology models. The models undergo a curation process that includes the assignment of MIRIAM identifiers to model components. It is recognised that this process requires great effort, and it is noted that the 17th release of the database (April 27th 2010) contains only 18950 annotated elements compared to a total number of model species and reactions in the database of over 82000. This lack of comprehensive annotation is claimed to be due to limited curator resources, and the paper itself states that the use of annotation tools, such as those developed from libAnnotationSBML, is among the resource's future development plans.

2.6.2 The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks

When used, MIRIAM annotations are commonly applied to models *after* their development, in order to facilitate their reuse. However, the use of semantic

annotations, and the linking to data resources that describe these terms, provided by libAnnotationSBML, can also be used in the development of metabolic networks. For example, the application of annotations referring to entries in metabolite databases can be used to automate the extraction of information such as chemical formulae, which is necessary in the reconstruction process. While libAnnotationSBML provides the facility for extracting such information, another layer of infrastructure is required above this to drive the reconstruction process.

Such an infrastructure has been developed and is described in the paper, *The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks* [103], published in the *Journal of Integrative Bioinformatics* (Publication 4, no ISI IF). The SuBliMinaL Toolbox was built to automate tasks that were often repeated during the iterative development of metabolic reconstructions in yeast and human, and forms a library of modules that can be used either individually, or as part of a workflow to generate a draft metabolic model from publicly available resources. The individual modules, and the workflow itself, are illustrated in Figure 6.

As noted in Section 2.4, both KEGG and MetaCyc provide the facility for exporting metabolic reconstructions. Many of the steps associated with generating a model from a reconstruction that is applicable to constraint-based analysis typically require manual curation. However, some of the steps are amenable to automation, providing the possibility of automating the process of generating a draft reconstruction for use in subsequent manual curation. It is the generation of models from KEGG and MetaCyc, along with the automation of subsequent curation steps, that the SuBliMinaL Toolbox attempts to address.

Modules in the workflow rely on these semantic annotations (and the libAnnotationSBML library) to extract relevant metadata from third party data resources. Such metadata includes SMILES [104] strings representing a metabolite, which can be used to determine a correct molecular formula and charge state for a given pH, and protein sequences or curated compartmentalisation data, allowing the inference of the intracellular compartment(s) in which a given metabolic reaction occurs.

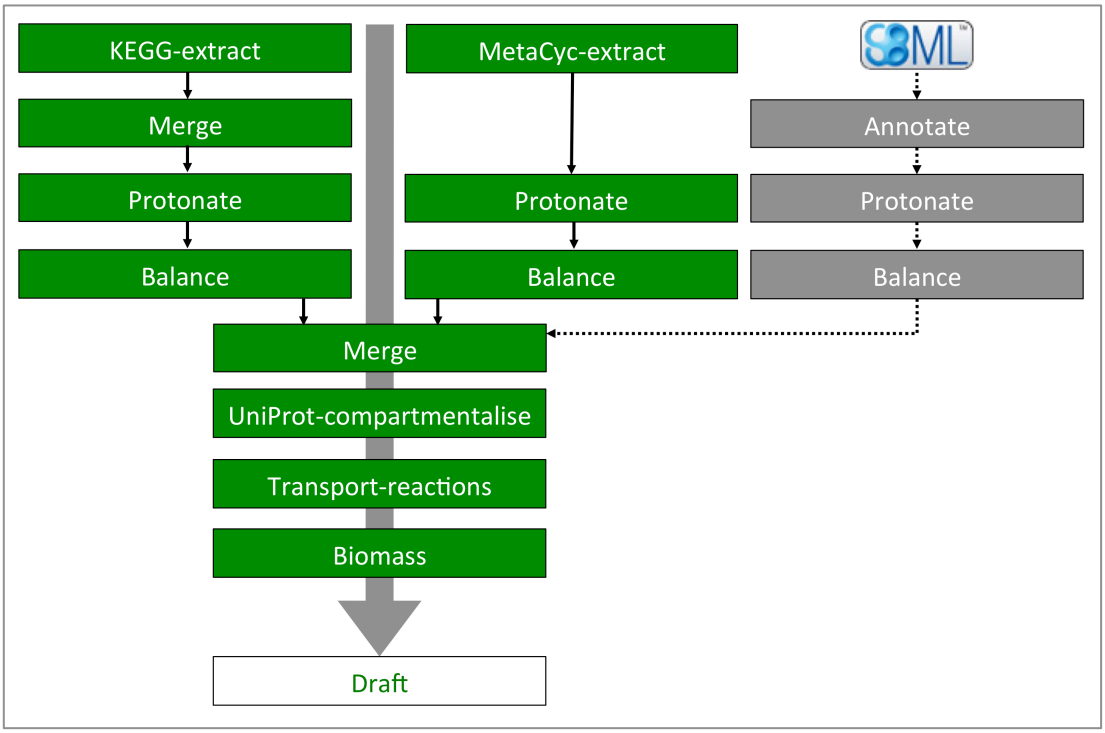

Figure 6: Illustration of how the individual modules of the SuBliMinaL Toolbox can be chained together to generate a draft metabolic reconstruction from publicly available data resources. The solid arrows indicate a workflow that extracts metabolic reactions and related data from the resources KEGG and MetaCyc, while the broken arrows illustrate how existing SBML models can be incorporated into such a pipeline.

The SuBliMinaL Toolbox therefore attempts to address some of the issues outlined by Poolman *et al.* [94] in developing metabolic reconstructions from public resources. A model of *Saccharomyces cerevisiae*, automatically generated from the workflow shown in Figure 6, was shown to be able to simulate biomass production from a minimal growth medium. The model itself consisted of an increase of 92% and 90% metabolites and metabolic reactions with respect to the manually generated version, Yeast 4.0. Despite this increase in size, the automated version showed comparable network connectivity to Yeast 4.0, with 79% of its metabolic reactions able to carry flux compared to 80% in the manually generated model. Nevertheless, models generated through automated means should still be considered to be drafts that require manual curation and validation in order to match the quality of other published reconstructions [89].

2.6.3 SBML Level 3 Package Proposal: Annotation

The current MIRIAM standard supports a subset of RDF to be used in the annotation of SBML models. While the advantages of applying MIRIAM formatted semantic annotations have been discussed, it has become clear to the SBML community that the current MIRIAM standard has a number of limitations. Specifically, these are:

- the inability to make statements about attributes. All MIRIAM annotations apply to the SBML element as a whole (e.g. a compartment, a species, a reaction). It is currently not possible to make statement about an SBML attribute (e.g. a compartment `volume` or a species `initialConcentration`);
- the inability to generate complex statements. Currently, the MIRIAM standard effectively supports "tagging" of SBML elements with identifiers from external data resources (see Figure 2). Complex statements, such as "protein X is modified by modifier Y in position X", cannot currently be stated;
- the inability to define relationships between statements. According to the current MIRIAM standard, if two or more statements are made, it is not clear whether both statements must apply (whether they form an "and" relationship) or whether the statements represent alternatives (an "or" relationship); and
- the inability to make negative statements.

The first three of these limitations have been addressed by Waltemath, Swainston *et al.*, in *SBML Level 3 Package Proposal: Annotation* [105], a proposal for an extension to the SBML language (Publication 5, no ISI IF). In the latest increment of SBML, Level 3, it has been recognised that the existing Level 2 SBML specification is insufficient to cover all of the needs of the systems biology modelling community. As such, it has been decided to retain a core SBML language specification, to which extension packages can be applied for more specialised applications. Examples of such extension packages include the Layout and Render packages, which define how an SBML model can be displayed graphically, the Flux Balance Constraints package, which extends the core SBML language to more fully support constraint-

based modelling approaches, and the Spatial Processes package, describing spatial modelling and the geometries of the spatial components involved. The Annotation package is another such proposal, which the SBML community has reviewed, and approved its development and inclusion in SBML Level 3.

The basic premise of the Annotation package is to remove the restrictions applied by the original SBML Level 2 specification determining how MIRIAM annotations are to be applied. In the Level 3 Annotation package, all features of RDF will be supported, allowing some of the limitations of the previous specification to be circumvented. For example, complex statements are supported through the use of RDF reification, a standard approach that allows statements to be made about statements. Protein modifications can be fully specified using this approach. Relationships between statements can now be specified, due to the full support of RDF containers and collections, allowing multiple statements to be defined as members of sets, ordered sets, closed sets or as alternatives to one another. The making of statements about attributes involves non-standard, yet valid, usage of RDF, in which XPath (http://www.w3.org/TR/xpath/) is used to define the SBML attribute about which an RDF statement is being made. However, negative statements cannot be made with RDF, but are supported by the OWL 2 Web Ontology Language, Version 2 (http://www.w3.org/TR/owl2-overview/). As the SBML Level 3 Annotation package is built upon RDF, the inability to make negative statements remains in this extension. The discussion regarding support of OWL 2, and therefore of negative statements, in the SBML Level 3 Annotation package is on going.

2.7 Conclusions

This work has sought to characterise the steps required in manually constructing metabolic reconstructions through community participation. This was approached by undertaking a number of projects in which genome-scale metabolic reconstructions were generated through development of a consensus between existing models. Such an approach was applied to reconstructions of *Saccharomyces cerevisiae* [73,85], *Salmonella typhimurium* [74] and *Homo sapiens* (*under review*). The key findings from applying such approaches were the requirements for

consolidating competing reconstructions in order to reduce duplication of effort across research groups, and for unambiguously identifying components within these reconstructions in order to promote their reuse and further development. These reconstructions and the publications describing them therefore represented the first examples of community development of metabolic reconstructions, and also the first reconstructions that adhere to community-developed data standards.

From the experience gained in developing such networks, techniques that automate many of the reconstruction steps were developed [98,103]. This work provided a framework for interpreting semantically annotated models, and also introduced an approach that supported the semi-automation of genome-scale metabolic reconstruction from existing data resources. The resulting software takes the form of a modular system that can be used in automated pipelines for the generation of draft networks, and also in curation and refinement of existing networks. These tools were applied to reconstructions of the three organisms outlined above, and are also be applied in on going work on a metabolic reconstruction of the human pathogen *Mycobacterium tuberculosis*, and to the reconstruction of the metabolism of a number of members of the cyanobacteria family.

Applying such a semi-automated approach provides the advantages of rapidly generating networks, allowing resources to be spent more productively on manual curation, and also provides a common format for draft reconstructions, upon which software to support the manual curation process can be developed.

3 Experimental data and kinetic modelling

3.1 Overview

Thus far, the modelling approach considered has been constraint-based stoichiometric modelling that does not consider the kinetics of the individual reactions within the network. While constraint-based modelling has many useful applications in modelling a system at steady state, it is only by considering reaction kinetics that a truly predictive *dynamic* model of an organism's metabolism can be generated. Kinetic models typically consist of a set of coupled ordinary differential equations (ODEs), each representing an individual metabolic reaction, which when solved by numerical integration predict the dynamic behaviour of the system [106].

A simple example of such an ODE can be taken from Michaelis-Menten kinetics [107], which can be modelled to provide the ODE defining the rate of formation of the product, P, where K_M can be interpreted as the binding affinity of the substrate S and enzyme E, and k_{cat} is the catalytic activity of the enzyme:

$$\frac{d[P]}{dt} = k_{cat}[E]\frac{[S]}{K_M + [S]} \qquad (1)$$

Clearly, such ODEs require parameterisation, and as such, kinetic modelling is reliant upon experimental data. As current publicly available data sets are frequently not applicable to the system under study, kinetic modelling can be thought of as being data poor. While work is being performed to infer kinetic constants from three-dimensional enzymatic structure and known parameters of homologous enzymes [108], there still remains a need to develop high-throughput techniques for experimental determination of such parameters [39].

Equation (1) illustrates that the rate of product formation ($d[P]/dt$) is a function of both the enzyme kinetic constants (k_{cat} and K_M) and the enzyme concentration, $[E]$. Some experimental strategies for determining $d[P]/dt$ follow an *in vivo* strategy, in which the determination of reaction rate is determined in the cell using fluorescence techniques [40]. This approach has some advantages in terms of performing the experiment in the actual environment in which the enzyme

operates, and therefore takes in to account enzyme activators and inhibitors that may be present in the system. However, this approach simply measures the maximum reaction rate, v_{max}, given in equation (2):

$$v_{max} = k_{cat}[E] \qquad (2)$$

Measuring v_{max} through *in vivo* studies therefore does not provide the granularity required to allow the kinetic constants and enzyme concentrations to be separated. In more advanced models, the enzyme concentrations themselves may be variables, given that they can vary due to transcriptional regulation. Furthermore, *in vivo* measurements of v_{max} capture the contribution of all isoenzymes into a single maximum reaction rate – the catalytic effect of each individual isoenzyme cannot be determined.

To remove these limitations, we have adopted an *in vitro* approach, which determines both kinetic constants and enzyme concentrations independently of one another, allowing the construction of more detailed kinetic models in which isoenzymes and variable initial enzyme concentrations may be considered. This approach therefore necessitates informatics support for both kinetic constant determination through *in vitro* assays, and quantitative proteomics approaches to determine enzyme concentrations.

3.2 Supporting enzyme kinetics assays for systems biology

The field of enzymology is mature and many kinetic constants have been determined, published and catalogued in databases such as SABIO-RK [42] and BRENDA [41]. While existing kinetic constants can be used in kinetic modelling [109], in many cases these values are not necessarily applicable to the system under study, being determined on different strains or organisms, or under experimental conditions with non-standard pH or temperature. Kinetic parameters are sensitive to these differing experimental conditions, and differ across orthologous enzymes across species. As such, the generation of accurate, predictive kinetic models is greatly facilitated by the generation of kinetic constants for the system under study,

following standardised experimental conditions that mimic the physiological conditions of the system [110].

The experimental approach that we follow is to perform absorbance assays on purified isoenzymes under investigation *in vitro*. The first step in this involves the overexpression and purification of individual isoenzymes, which provide an isolated, pure isoenzyme to be used in assays. The assays produce a number of time-series plots, indicating the rate of product formation at different initial substrate concentrations. From this raw data, kinetic constants defining the reaction can be determined.

From an informatics perspective, the requirements to support such kinetic assay experiments are therefore:

- to support the acquisition of sufficient metadata to describe the experiment with enough detail to allow for the reuse of the experimental data;
- to apply fitting algorithms to analyse raw experimental data to determine secondary, kinetic constants;
- to support the storage and dissemination of the original raw experimental data, allowing the raw data to be reanalysed in future as fitting algorithms develop;
- to support the storage and dissemination of the derived kinetic parameters, in such a way that kinetic models may be parameterised automatically; and
- to adhere to existing data standards where possible.

Fulfilment of these requirements required an initial appraisal of existing software infrastructure that supports enzyme kinetics data. A number of tools existed that could determine kinetic parameters from experimental time-series data, including GraFit (Erithacus Software Ltd., Horley, UK) and COPASI [111]. While both tools could be used to perform the analysis, neither offered support for metadata capture or data storage and dissemination. The data resources BRENDA and SABIO-RK provided the facility for storing kinetic parameters, but did not allow the original raw experimental data to be archived, nor was there an existing data standard to

provide this functionality. The archiving of the original raw experimental data was considered to be important to allow kinetic parameters to be associated back to the original data sets from which they were derived, providing additional confidence in the validity of such parameters. In addition, raw data storage provides the facility for reanalysis of the data to be performed as data analysis algorithms increase in sophistication. Considering the lack existing tools that satisfied the above requirements, it was decided to develop a new suite of tools, but to reuse existing resources where possible.

Regarding data storage and the requirement to follow existing standards, both SABIO-RK and BRENDA adhere to the Standards for Reporting Enzymology Data (STRENDA) Guidelines (http://www.beilstein-institut.de/en/projekte/strenda/guidelines/). The STENDA Guidelines describe the minimum set of metadata that must be supplied with enzyme kinetic data to promote its reuse, and cover description of the experiment, description of enzyme activity data, and definitions of experimental conditions. Adherence to the STRENDA Guidelines when publishing kinetic data has now been recommended by a number of journals [112]. As such, a further requirement for the enzyme kinetics infrastructure was the implementation of the STRENDA Guidelines.

SABIO-RK was selected to act as the repository in which derived kinetic parameters would be stored. At the time of development, SABIO-RK provided more support for semantically annotated data than BRENDA, including the facility to export kinetic parameters in MIRIAM-standardised SBML files. The storage and dissemination of such parameters in a format that matches those in the kinetic models themselves greatly facilitates the automated parameterisation of such models [113].

The storage of raw data is performed by a newly developed database, MeMo-RK, which stores raw data and metadata in parallel to storage of derived kinetic constants in SABIO-RK. MeMo-RK makes this data accessible via both a web browser and web service interface, and provides a link between itself and SABIO-RK, allowing navigation from raw data to secondary data and vice versa.

To fulfil the requirements of the enzyme kinetics infrastructure, a tool, the KineticsWizard, was developed to interface with the instrument software, prompt

the user for sufficient metadata to satisfy the STRENDA Guidelines, determine kinetic parameters from raw data, and submit both the experimental raw data and kinetic parameters to MeMo-RK and SABIO-RK respectively for storage and dissemination. Most of the functionality of the Wizard is instrument independent, and can be used with time series data generated on any instrument (see Figure 7).

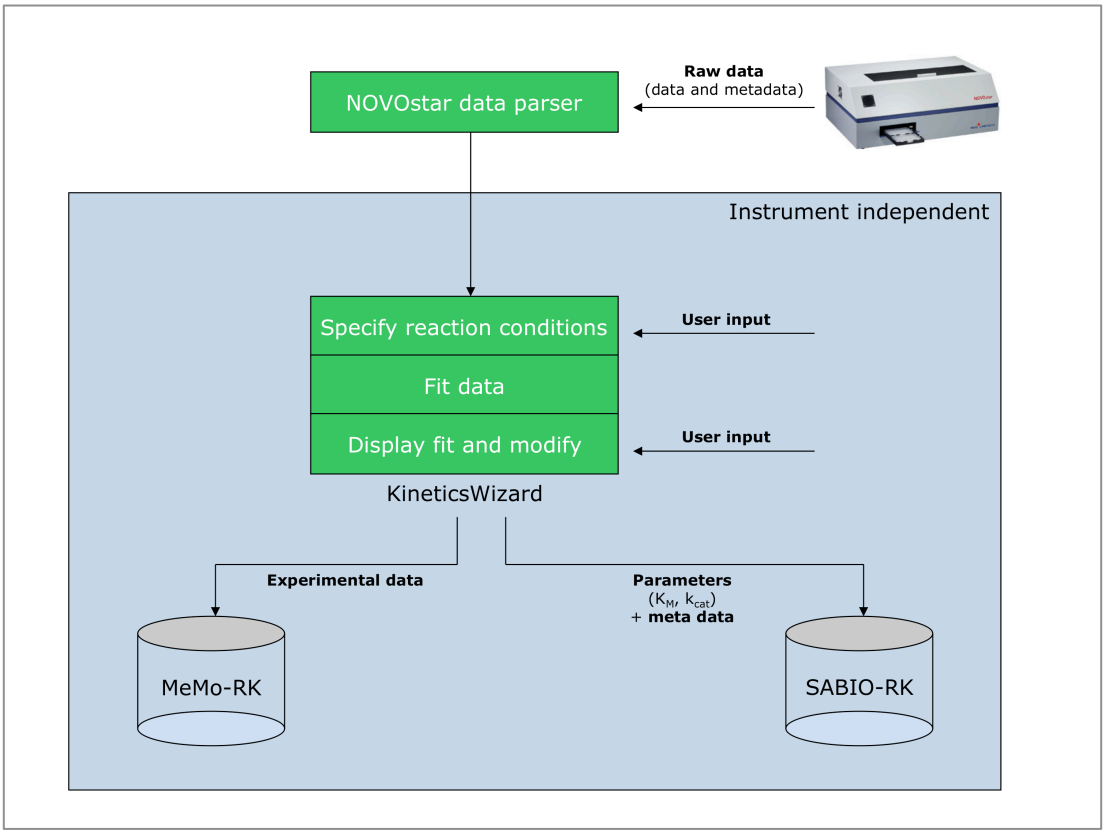

Figure 7: Data flow through the KineticsWizard. Experimental data is extracted from instrument software, metadata is collected from the user, and kinetic constants derived from the experimental data by automated fitting. The fit is displayed to the user and can be modified by hand if required. Experimental raw data and derived kinetic parameters can then be submitted to the databases MeMo-RK and SABIO-RK respectively.

An important consideration of the KineticsWizard is in its definition of reactions and kinetic parameters using third-party database identifiers, such as KEGG reaction ids, ChEBI terms and Systems Biology Ontology (SBO) terms. Specifying experimental metadata in such terms – using the same identifiers as those used in the consensus model described in Section 2.3 – allows for the automated parameterisation of such models with this experimental data (see Section 3.4).

This work gave rise to a publication in *FEBS Journal*, *Enzyme kinetics informatics: from instrument to browser* [114] (Publication 6, ISI IF 3.1). This paper describes the first integrated solution for supporting the extraction, analysis, storage and dissemination of enzyme kinetics data, and the first example of software that implements the STRENDA Guidelines and automates the process of submitting kinetics data to a publicly available repository. The system is used by experimentalists in the MCISB and has been cited eight times including work discussing the automation of kinetic modelling in systems biology [115,116].

A limitation of the current implementation is that its fitting algorithm assumes that the reaction follows Michaelis-Menten kinetics. While this assumption holds for a number of reactions, many other reaction mechanisms are known. Future developments of the KineticsWizard will therefore consider a range of kinetic mechanisms. Doing so will involve the application of more sophisticated parameter fitting algorithms [117], and will also necessitate the expansion of the user interface to allow the specification of more complex reaction mechanisms.

3.3 Quantitative proteomics in kinetic modelling

The kinetic modelling approach that we follow is dependent upon the determination of *in vivo* enzyme concentrations, which act as parameters in kinetic models. The determination of protein concentrations, known as quantitative proteomics, can follow a number of approaches including the use of mass spectrometry [118], and this is the approach that is considered here.

Mass spectrometry (MS) determines the mass-to-charge ratio of charged particles. In a typical qualitative proteomics experiment, a protein sample is digested to form peptides, which are then introduced to a mass spectrometer. The peptides are then charged and separated according to their mass-to-charge ratio. From this collection of measured mass-to-charge ratios, algorithms can be used to determine the identity of the peptides within the sample by comparing these measured masses against masses determined from an *in silico* digestion of protein sequence databases [119]. This basic technique has subsequently been developed and is now able to identify thousands of proteins in a single experiment [120].

Mass spectrometry, however, is not in itself a quantitative method. Signal intensities detected across different peptides are not proportional to the relative concentrations of each peptide, due to a number of factors including inconsistent protein digestion, chromatographic separation and ionisation efficiency. However, signals detected for physiochemically identical peptides *are* proportional to their relative concentrations, and this behaviour can be exploited in quantitative proteomics. Mass spectrometry approaches to quantitative proteomics therefore rely on isotopic labelling of samples. By isotopically labelling a given sample, a labelled and non-labelled peptide in two samples shares identical physiochemical properties, and therefore behaves consistently during the sample preparation, chromatography and ionisation processes, providing comparable peak signals intensities. However, as these peptides differ in mass, due to the presence of an isotopic label, they can be distinguished by mass spectrometry [121].

One such quantitative approach is isobaric tag for relative and absolute quantitation (iTRAQ) [122], which allows for isotopic labelling and therefore relative quantitation determination of up to eight samples simultaneously. This technique was utilised in a large-scale study of growth of *Saccharomyces cerevisiae* under different conditions [123], and the data from this study was used in a pilot study for data management of quantitative proteomics data.

The data itself consisted of a number of MS files consisting of multiplexed samples labelled with iTRAQ tags, and generated under different growth rates and nutrient limitations. Requirements for analysis and dissemination of this data were:

- to identify peptides (and therefore proteins) within the sample, using an existing database search algorithm;
- to determine relative quantitation of each peptide (and therefore protein) within the pooled sample from iTRAQ reporter ions;
- to apply metadata to the MS data, describing the experimental conditions under which the individual samples within the data were acquired; and
- to format the metadata, MS data, peptide identifications and quantitations to comply with an existing data standard, to allow for its dissemination and reuse.

Again, existing software and standards were to be used, with the goal of exploiting these resources in a single workflow that fulfils the above requirements. The data format selected for storage and dissemination of the integrated data was PRIDE (PRoteomics IDEntification database) XML [124], a community developed standard that allowed for much of the necessary data to be captured. The use of standardised data formats and the submission of data to public repositories is becoming increasingly prevalent, and has been cited as being a future prerequisite for publication of proteomics articles in journals such as *Nature Biotechnology* [125].

At the time of development, no integrated solution existed to perform the above tasks. Indeed, there was only one software resource supporting the capture of proteomics data in PRIDE format, Proteome Harvest [126], a tool that relies heavily on manual input. However, a number of database search engines were available to perform peptide and protein identification, and quantification algorithms existed for analysis of iTRAQ data. A tool was therefore developed, the PrideWizard, to integrate the existing resources and automate the process of moving from raw experimental data to a standardised representation of the derived protein identifications and quantifications.

The PrideWizard allows the import of MS data, which is then integrated with peptide/protein identifications from the database search engine, Mascot [127]. Metadata is collected, describing the experiment and the individual pooled samples, a process that exploits the EBI Ontology Lookup Service [128], to allow the specification of metadata in standardised, controlled vocabulary terms. The PrideWizard then generates relative quantitations for each peptide by calling the iTracker quantitation software [129], before integrating these results with the peptide / protein identifications and metadata to generate a PRIDE XML formatted output file, which can then be uploaded to the publicly available PRIDE Database.

A major challenge in the representation of quantitative proteomics data in a common format was the fact that, at the time of development, no formats existed for the mark-up of quantitative data. It was known that the Human Proteome Organisation Proteomics Standards Initiative (HUPO PSI) were developing a number of community-driven standards for representing quantitative proteomics data,

including mzIdentML and mzQuantML for identifications and quantitations respectively. However, mzIdentML was only released in August 2009 [130], and work on mzQuantML is still on going. As such, the development of the PrideWizard necessitated the extension of the PRIDE XML format to allow for the capture of quantitative data. PRIDE XML provides a facility for such an extension, allowing user-definable terms to be applied to any existing element in the PRIDE XML structure. The specification of appropriate user-definable terms for the mark-up of quantitative data was performed in consultation with members of the HUPO PSI community. This involved the extension of the PRIDE ontology to include terms specific to the iTRAQ approach.

This work resulted in a publication in *BMC Proteome Science*, *An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ* [131] (Publication 7, ISI IF 2.5). This paper has been cited 21 times, in manuscripts discussing proteomics data standards [126,132,133] and iTRAQ data analysis [134,135]. Perhaps the most significant citation was in the article *PRIDE: new developments and new datasets* [136], which highlighted the successful collaborative effort between members of the HUPO PSI community and the developers of the PrideWizard in extending the PRIDE XML format for quantitative proteomics data.

Due to the development of the PrideWizard, the proteomics data generated in the *Saccharomyces cerevisiae* growth pilot project represented the first example of iTRAQ data submitted to the PRIDE data repository. As of August 2011, the publicly available PRIDE Database contains 19 data sets that utilise the PRIDE XML iTRAQ extension in addition to those generated in the *Saccharomyces cerevisiae* growth pilot study described here.

While the pilot project concentrated on determination of *relative* quantitations of proteins, the kinetic modelling approach described in Section 3.1 is dependent upon *absolute* measurement of enzyme concentrations. This is performed by following the recently developed QconCAT [46] approach, which relies upon the generation of heavily labelled marker peptides that act as unique references for a given set of proteins. These peptides are concatenated into a single protein (a QconCAT protein) that is expressed and grown on heavily labelled media, such that each arginine and

lysine residue is labelled with six ^{13}C atoms. By mixing and digesting a known quantity of this labelled protein with the sample under investigation, pairs of heavy and light peptides can be detected by mass spectrometry and their relative concentration determined by measuring the ratio of their peak intensities. As the absolute concentration of the heavily labelled peptide is known, the absolute concentration of the sample peptide (and therefore the protein from which it originated) can be determined.

We follow this approach in order to quantify enzymes necessary for the parameterisation of a given kinetic model. The starting point for the workflow is the unparameterised model itself. An *in silico* digest of the amino acid sequence of each metabolic enzyme is performed, producing the set of peptides that the protein is likely to give rise to. From this set, peptides that can be expected to act as unique, proteotypic "marker" peptides for the protein can be selected, and one or more of these can be incorporated into the QconCAT protein to be used in quantitation of all enzymes in the model. Brownridge *et al*. [137] discusses the selection of suitable marker peptides for a given protein, and while our work entailed manual selection of marker peptides, attempts have been made to automate this process [138,139].

The requirements for data analysis and dissemination of the absolute quantitative data were the same as for the relative quantitative data generated in the *Saccharomyces cerevisiae* growth pilot project following the iTRAQ approach. Similar to the situation with iTRAQ data, at the time of development there existed no dedicated informatics software that fulfilled these requirements for QconCAT data. The approach taken was therefore to extend the existing PrideWizard to deal with the more specialised case of absolute quantitative proteomics following the QconCAT strategy.

A pipeline was therefore established in order to automate the steps of converting raw MS data of QconCAT samples to a standardised representation of peptide / protein identifications and quantitations, and appropriate metadata, that could be disseminated and used in kinetic model parameterisation (see Figure 8).

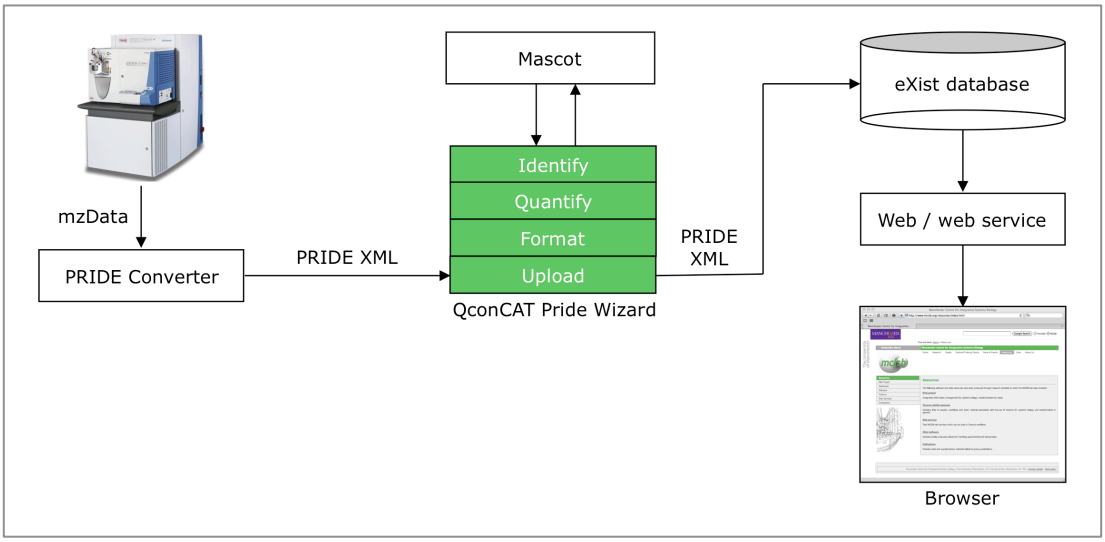

Figure 8: Workflow for the analysis and dissemination of absolute quantitative proteomics data using the QconCAT methodology.

Raw MS data is exported from the instrument and metadata applied through use of the PRIDE Converter [140], a publicly available tool generated at the European Bioinformatics Institute (EBI). The annotated MS data is passed to the QconCAT PrideWizard, an extended version of the original PrideWizard, which performs the integration of identifications and quantitations. The quantification of QconCAT fundamentally involves two steps: i) identification of heavy / light peptide pairs; and ii) determination of heavy / light peptide pair peak intensity ratios. Identification of peptide pairs was performed through a Mascot database search against the UniProt database, providing protein identifiers that match those used to annotate the kinetic models to be parameterised. Once identified, a port of the existing SILACAnalyzer quantitation algorithm [141] quantifies peptide pairs. The identifications and quantitations are merged into a PRIDE XML document, which is then automatically uploaded to a native XML database, eXist (http://exist.sourceforge.net/), upon which a web browser and web service interface has been developed. The web service interface provides access to all of the data stored in the PRIDE XML document, and a specialised interface has been written to automatically parameterise kinetic models encoded in SBML.

A requirement for the dissemination of the absolute quantitative proteomics data was to provide the facility for viewing both the derived, secondary data in terms of the calculated protein concentrations, and the original raw data from which these

concentrations were derived. Providing access to the original raw data provides additional confidence to the consumer of the derived protein concentrations, and can be used in analysis of discrepancies between these values and those predicted by mathematical modelling. While PRIDE XML provides the facility for storing mass spectra, extracted ion chromatograms, displaying the peak areas of the heavy and light peptides, could not be represented. Support for storage of chromatograms is implemented in the mzML data format [142], but the PRIDE XML format does not as yet formally support this. As such, as an interim solution, the PRIDE XML format was extended to allow for the mark-up of extracted ion chromatograms within the current PRIDE XML structure, with additional user-defined terms added to map these chromatograms to the individual peptides from which they arise.

The final pipeline therefore produced a solution for the analysis and dissemination of absolute quantitative proteomics data, from the raw instrument data through to visualisation of identifications, quantifications and raw data in a web browser. In addition, a web service interface has been developed that can be accessed to automate the process of parameterising kinetic models. Furthermore, the data is generated in PRIDE XML format, and can therefore be uploaded to the centralised PRIDE Database for public dissemination. This work was described in the paper, *A QconCAT informatics pipeline for the analysis, visualization and sharing of absolute quantitative proteomics data* [143], which was published in *Proteomics* (Publication 8, ISI IF 4.8) and has been cited twice, including an article reviewing existing approaches in quantitative proteomics [144].

3.4 Automation of the generation of kinetic models from genome-scale metabolic networks

The ultimate goal of supporting both enzyme kinetics assays and quantitative proteomics experiments is to use data generated through such approaches in the parameterisation of kinetic models. The kinetic models are themselves generated from the larger, genome-scale metabolic networks. As such, the work described thus far (genome-scale network reconstruction, support of kinetic assays and quantitative proteomics experimentation) represents individual components of a larger data-integration pipeline, in which the challenge is to automate the

generation of kinetic models from genome-scale reconstructions and experimental data. The data sets involved in such integration are illustrated in Figure 9.

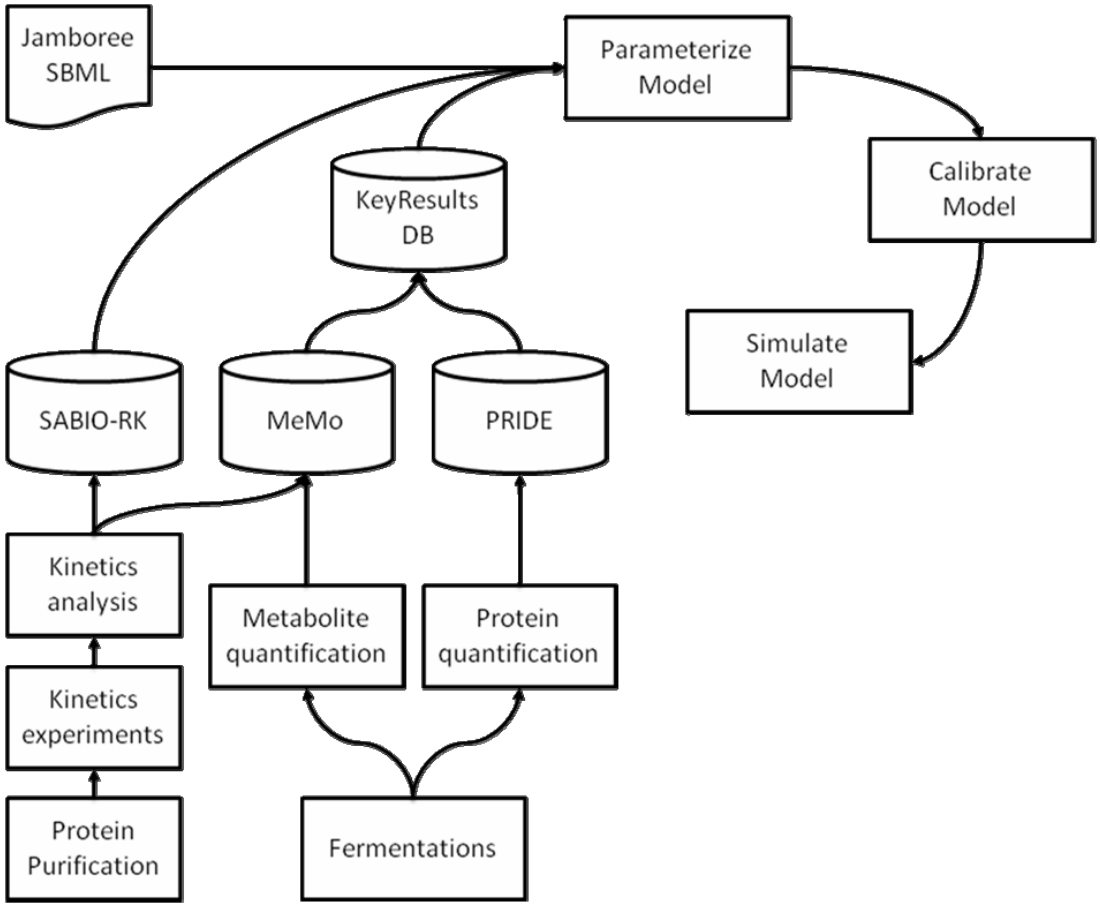

Figure 9: Illustration of data integrated in the development of kinetic models. The data resources SABIO-RK [42], MeMo [150] and PRIDE [124] are used to store enzyme kinetic constants, quantitative metabolomics and quantitative proteomics data respectively. The quantitative omics data sets can be held in an instance of the KeyResults Database (Jameson *et al.*, unpublished) before being integrated with a genome-scale metabolic reconstruction (denoted "Jamboree SBML") in a model parameterisation step. The parameterised model can then be calibrated and simulated in software tools such as COPASI. This illustration is taken from Swainston *et al*. [145], which describes the integrative pipeline followed in generating kinetic models from experimental data.

As experimental work is expensive, and as large-scale kinetic modelling is challenging, kinetic modelling is typically performed on small models representing a given metabolic pathway. The question remains, then, of which pathways should be

selected for kinetic modelling from a genome-scale metabolic network. The approach that we follow is to prioritise carbon consumption, as the extraction of energy from carbon sources is central to many life forms and is sufficiently well studied to act as a test case for our systems biology approach. As such, flux balance analysis was performed on the genome-scale metabolic model, Yeast 4.0 [85], to determine the metabolic pathways that carry the greatest carbon flux. From this analysis, individual pathways were exported from the genome-scale model, and these then acted as a requirements specification for the subsequent experimental work, defining the set of metabolic reactions that require parameterisation.

Upon completion of the experimental work, appropriate repositories are populated with derived, secondary values that can be used to parameterise the kinetic model. As both the experimental data and the model are annotated with consistent identifiers, and the repositories themselves are equipped with web service interfaces, the task of parameterising models can be automated.

The automation of model parameterisation is discussed in *Integrative Information Management for Systems Biology* [145], published in *Lecture Notes in Computer Science* (Publication 9, no ISI IF). The paper has been cited twice, including a review paper discussing the use of workflows in the biological sciences [146]. This work describes the entire pipeline utilised in generating kinetic models, from analysis and storage of raw experimental data, through to the workflows that exploit this data in the generation, parameterisation, and simulation of kinetic models through the COPASI web service interface [147]. A subsequent manuscript by Li *et al*. describes this collection of workflows in detail [116], and the use of the workflow engine Taverna [148] in their execution. The work of Li *et al*. was supplemented by a commentary article [149], which described the work as “a promising evolutionary step forward in the development of techniques that facilitate the construction, analysis and dissemination of computational models”.

3.5 Conclusions

The experimental data and kinetic modelling aspect of this work resulted in the development of computational techniques and infrastructure to support the generation of parameters required in kinetic modelling from experimental data.

Typically, experimental data acquisition and management is performed independently of any consideration of subsequent use in systems biology modelling, with kinetic models typically being built manually from existing experimental data sets [40]. However, due to the increase in complexity of systems biology modelling, and the wide range of diverse datasets required for model construction, a recent poll found that data and tool integration was the single most important development sought by systems biology researchers [151]. The work presented here contributes to this requirement, providing the first example of support for experimental data capture, analysis and dissemination with the specific goal of automating the integration of such data into kinetic models.

4 Conclusion

4.1 Future requirements for experimental data management

Although it has been shown that the development of computational infrastructure can facilitate metabolic systems biology research, at both the genome and pathway scale, a large number of challenges remain.

From an experimental data management perspective, the field of metabolomics lags those of proteomics and transcriptomics in terms of the informatics support available. Metabolomics is a key experimental technique in systems biology, providing measured metabolome values to validate those predicted by modelling. However, it is noticeable that at the current time, there is little dedicated support or a widely used publicly available repository for metabolomics data. The metabolomics database MeMo [150] was developed to address this need, but does not appear to have been taken up by the metabolomics community. Furthermore, while a Metabolomics Standards Initiative (MSI) exists, their website suggests little activity since 2007 [152]; a situation which contrasts sharply with the vibrancy of the equivalent community in proteomics, the HUPO PSI. It is hoped that this situation will improve with the introduction of MetaboLights, which has been described as the "missing metabolomics repository" and is being developed at the European Bioinformatics Institute (EBI) (http://metabolomes.sourceforge.net/). MetaboLights is expected to act as a centralised repository for experimental metabolomics data in the same way that PRIDE [124] and ArrayExpress [153] fulfil these roles at the EBI for proteomics and transcriptomics data respectively, and will hopefully provide the impetus required to promote the sharing of metabolomics data.

However, it is not just the storage of metabolomics data that is still limited: metabolite identification still represents a major problem [154]. While both proteomics and metabolomics can be performed by mass spectrometry, identification of metabolites from mass spectrometry data represents a far greater challenge than of peptides. All peptides share largely the same physiochemical properties, and as such, under tandem mass spectrometry, peptides fragment in a relatively predictable manner, allowing the probabilistic matching of measured

fragment spectra against *in silico* predictions of fragment spectra generated from peptide sequences. Metabolites, however, are much more heterogeneous, and as such, their fragmentation patterns cannot yet be predicted accurately. Therefore, while the field of proteomics has given rise to a number of database search engines such as Mascot [127], Sequest [155] and X! Tandem [156], metabolomics still lacks a robust means of automatically identifying metabolites from mass spectrometry data.

As metabolomics is of such importance to metabolic systems biology, the continued development of informatics infrastructure to support the technique – both in terms of data analysis and dissemination – is likely be a requirement of the informatics community over the coming years.

Another analytical technique that is of direct relevance to metabolic systems biology is fluxomics [157]. Fluxomics provides measurement of intracellular metabolic fluxes by isotopically labelling a growth media nutrient, for example ^{13}C-glucose, and applies mass spectrometry to quantify the distribution of this labelled atom in the endometabolome. From this distribution, intracellular metabolic fluxes can be inferred. Such data can be used as a constraint in genome-scale modelling techniques such as flux balance analysis (FBA) [158], and can be used to validate kinetic modelling predictions (Malys *et al*., unpublished work). While software support exists for both the analysis [159] and visualisation [160] of fluxomics data, there appears to be no plans for developing a standard to support its sharing and dissemination.

This lack of publicly available fluxomics data is disappointing given its applicability to constraint-based modelling approaches in particular. FBA predicts intracellular fluxes, and therefore these predictions can either be validated by comparison with experimentally measured fluxomics data, or the data itself can be used to improve predictions by constraining the search space of viable intracellular fluxes [158]. Given the utility of fluxomics data, increased computational support for both its analysis and dissemination would be a welcome contribution to the field of metabolic systems biology.

Constraint-based modelling has an advantage over kinetic modelling in that it is not necessarily dependent upon experimental data: constraint-based modelling can be performed solely on the stoichiometric network provided by genome-scale models and, in its simplest form, requires no further parameterisation. However, it is being seen that constraint-based modelling has its limitations. A clear limitation is its reliance upon an objective function, which in the case of microorganisms has typically been the assumption that the cell is optimised for growth. For more complex, multicellular organisms, this assumption may no longer hold, and the objective of a cell in one tissue is likely to differ greatly from that in another. As such, the overall assumption of an objective function may not be applicable to more complex systems, and instead, the behaviour of the organism may be better elucidated through observation. Constraint-based modelling is therefore, like kinetic modelling, becoming more dependent on experimental data. An example of recent work that incorporates experimental measurements into constraint based modelling techniques is the metabolic analysis tool, iMAT [161]. iMAT implements the method of Shlomi *et al*. [162] of integrating measured transcriptomics and proteomics data with genome-scale reconstructions in order to predict intracellular metabolic fluxes. This method is built on the observation that there is correspondence between gene / protein expression and measured [163] and predicted [164] metabolic flux through reactions catalysed by the expressed enzymes. As such, experimentally measured omics data can be used as a guide to predicting the intracellular metabolic state, and this approach has been applied to the generation of tissue-specific models [165] from a generic metabolic reconstruction of human [166].

Systems that systems biology attempts to model are therefore increasing in complexity, and related to this is an ever-expanding dependency of modelling techniques on experimental data. Development of informatics infrastructure to support the management of such data, such as those approaches described in this thesis, is consequently likely to grow in importance over the coming years.

4.2 Future requirements for model development and management

As well as benefitting from improved informatics in processing and storing experimental data, systems biology also has a requirement for improved model management infrastructure. Some challenges in sharing models have been overcome, as the widespread use and support of the model representation language, SBML, demonstrates. Efforts to increase the reuse of such models through of application of semantic annotations following the MIRIAM standard are also welcome, although the use of such annotations is still not standard practice. Centralised repositories, such as the BioModels Database and JWS Online [167], in which published models are validated before being shared with the community are also useful additions to the systems biology infrastructure. However, such resources only store the final, published model and do not support the iterative model generation process. To do so, further development of tools that can be used in incremental model building, such as MEMOSys [168], rBioNet [169] and the SuBliMinaL Toolbox [103], will be required.

Another requirement is that of model merging, which is provided by both the SuBliMinaL Toolbox and semanticSBML [101], and will become increasingly important as larger scale, composite kinetic models are built from existing smaller models of individual pathways. Visualisation of models is also an issue, and the systems biology community have recognised this with the development of the System Biology Graphical Notation (SBGN) [170], which is hoped to standardise the representation of biochemical knowledge in the same way that circuit diagrams provide a universally understood representation of electronic systems. Software support for this format is already provided in tools, such as CellDesigner [171], Payao [172] and Arcadia [173], but it is noted that rendering of large (genome-scale) networks is still problematic with existing software.

In addition to supporting the provision of models in a common format, the systems biology community has also been active in the development of a mark-up language for specifying the actions to be performed upon them in the form of the Simulation Experiment Description Markup Language (SED-ML) [174]. SED-ML implements the Minimum Information About a Simulation Experiment (MIASE) guidelines [175], and provides an unambiguous description of the simulation performed upon a given

model. Specifying the model, the simulation, and the simulation result in a common format, such as the recently introduced Systems Biology Results Markup Language (SBRML) [176], provides a clear audit trail of the modelling process, and facilitates comparison of both simulation algorithms [177] and competing models of the same biological system.

While the generation of standards is a useful and necessary step for the systems biology community to take, recently introduced standards such as SED-ML and SBRML are yet to be supported by any software implementation. A further issue is that existing tools that support SBML tend to focus upon a single function, be that model construction, visualisation, constraint-based analysis, etc. While the SBML standard is widely used, and endeavours such as the Systems Biology Workbench [178] facilitate the task of working across a number of software tools, the field of systems biology informatics is still rather fragmented. The situation is limited further by the fact that certain packages do not fully support recent versions of the SBML specification, or have implemented their own "dialect" of SBML. As a consequence of this, models generated with a given tool frequently cannot be used with a separate tool, despite both claiming to fully support the SBML standard. While this fragmentation of software infrastructure is perhaps inevitable, given the range of expertise and interests that individual research groups possess, efforts should be made to encourage full interoperability of models between software tools. This process has been greatly helped by the development of programming libraries supporting a range of programming environments, such as libSBML [95], and it is encouraging to learn of related efforts with the development of libraries supporting newly introduced standards such as libSBGN (http://libsbgn.sourceforge.net/), libSEDML (http://libsedml.sourceforge.net/) and libSBRML (Dada *et al*., unpublished).

4.3 Contributions of this work to data integration in metabolic systems biology

In addition to software and standard representations to support the exchange of models and experimental data, there is also a need to develop infrastructure to integrate the two. The work that supports this thesis attempts to address some of the issues relevant to data integration in systems biology. This work includes the

first example of a community developed genome-scale metabolic reconstruction that provides full semantic annotation, which acts as a template for subsequent modelling efforts that involve automated parameterisation with experimental data (Publication 1). This reconstruction was subsequently expanded to generate a model that could be used in constraint-based analysis (Publication 2). Building on the experiences gained from producing such models, a programming library and a toolkit to support subsequent semi-automated development of genome-scale metabolic reconstructions for any organism were introduced (Publications 3 and 4). Recognising the importance of standardised semantic annotations of models, an extension to the existing SBML annotation format has been proposed and accepted by the SBML community (Publication 5). The importance of developing models in which components are unambiguously identified has been discussed both in the context of sharing and reusing models, and in their increased amenability to experimental data integration.

On the experimental side, an existing standard representation for proteomics data has been extended to allow for the representation of quantitative data (Publication 7), and extended in the development of an automated pipeline for the analysis, storage and dissemination of absolute quantitative proteomics data (Publication 8). Similar work was performed for enzyme kinetic data, with the development of an automated pipeline for spectrophotometric data, which introduced the first system for storage and dissemination of raw experimental data, and the first solution that automates the submission of kinetic parameters to a publicly available resource (Publication 6). Given the increasing dependence of modelling approaches on experimental data, and the fact that acquisition of such data continues to be expensive and still reliant upon highly skilled experimentalists, techniques considered in this thesis, that facilitate the analysis and reuse of experimental data, are likely to grow in importance.

These contributions were further integrated into a workflow for automating the generation and simulation of kinetic models from genome-scale reconstructions and experimental data from enzyme kinetics and quantitative proteomics approaches (Publication 9). Automation of the systems biology pipeline, from experimental data through to model simulation, is currently not a widespread

practice. However, with improved instrumentation allowing for the ever more rapid generation of experimental data, and increasing complexity of models themselves, both in terms of their size and their development across cellular functions such as metabolism and cell signalling [79], such automated approaches will become increasingly necessary over the coming years. It is hoped that the work presented here to support this thesis forms a contribution towards this effort, and that this work will be subsequently developed further to fully realise the potential of systems biology.

5 References

[1] Characterization of the yeast transcriptome. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW. *Cell*. 1997, **88**, 243-51.

[2] Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL. *Biotechnol Genet Eng Rev*. 1996, **13**, 19-50.

[3] Systematic functional analysis of the yeast genome. Oliver SG, Winson MK, Kell DB, Baganz F. *Trends Biotechnol*. 1998, **16**, 373-8.

[4] Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. Van Regenmortel MH. *EMBO Rep*. 2004, **5**, 1016-20.

[5] Our genome unveiled. Baltimore D. *Nature*. 2001, **409**, 814-6.

[6] Initial sequencing and analysis of the human genome. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker

ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. *Nature*. 2001, **409**, 860-921.

[7] Systems biology: a brief overview. Kitano H. *Science*. 2002, **295**, 1662-4.

[8] A new approach to decoding life: systems biology. Ideker T, Galitski T, Hood L. *Annu Rev Genomics Hum Genet*. 2001, **2**, 343-72.

[9] Cybernetics or control and communication in the animal and the machine. Wiener N. *MIT Press, Cambridge, MA*. 1948.

[10] Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Kell DB, Oliver SG. *Bioessays*. 2004, **26**, 99-105.

[11] Modeling the heart--from genes to cells to the whole organ. Noble D. *Science*. 2002, **295**, 1678-82.

[12] Regulation of gene expression by a metabolic enzyme. Hall DA, Zhu H, Zhu X, Royce T, Gerstein M, Snyder M. *Science*. 2004, **306**, 482-4.

[13] Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Herrgård MJ, Lee BS, Portnoy V, Palsson BØ. *Genome Res*. 2006, **16**, 627-35.

[14] Biochemistry. 5th edition. Berg JM, Tymoczko JL, Stryer L. *New York: W H Freeman*, 2002.

[15] The history of the word 'metabolism'. Bing FC. *J Hist Med Allied Sci*., 1971, **26**, 158-80.

[16] Metabolism Is Not Boring. Ray LB. *Science*. 2010, **330**, 1337.

[17] Metabolism of tumours. Warburg O. *Biochem Zeitschr*. 1923, **142**, 317-33.

[18] Targeting metabolic transformation for cancer therapy. Tennant DA, Durán RV, Gottlieb E. *Nat Rev Cancer*. 2010, **10**, 267-77.

[19] Dyslipidemia and glucose dysregulation in overweight and obese patients. LeRoith D. *Clin Cornerstone*. 2007, **8**, 38-52.

[20] Changes in network activity with the progression of Parkinson's disease. Huang C, Tang C, Feigin A, Lesser M, Ma Y, Pourfar M, Dhawan V, Eidelberg D. *Brain*. 2007, **130**, 1834-46.

[21] Systems biology, metabolic modelling and metabolomics in drug discovery and development. Kell DB. *Drug Discov Today*. 2006, **11**, 1085-92.

[22] Understanding human metabolic physiology: a genome-to-systems approach. Mo ML, Palsson BØ. *Trends Biotechnol*. 2009, **27**, 37-44.

[23] Manufacturing molecules through metabolic engineering. Keasling JD. *Science*. 2010, **330**, 1355-8.

[24] Synthetic biology: new engineering rules for an emerging discipline. Andrianantoandro E, Basu S, Karig DK, Weiss R. *Mol Syst Biol*. 2006, **2**, 2006.0028.

[25] Group contribution method for thermodynamic analysis of complex metabolic networks. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. *Biophys J*. 2008, **95**, 1487-99.

[26] Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. Fleming RM, Thiele I, Nasheuer HP. *Biophys Chem*. 2009, **145**, 47-56.

[27] von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. Fleming RM, Thiele I. *Bioinformatics*. 2011, **27**, 142-3.

[28] Thirteen years of building constraint-based in silico models of Escherichia coli. Reed JL, Palsson BØ. *J Bacteriol*. 2003, **185**, 2692-9.

[29] Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, Baart G, Schwartz JM. *BMC Syst Biol*. 2010, **4**, 114.

[30] KEGG: kyoto encyclopedia of genes and genomes. Kanehisa M, Goto S. *Nucleic Acids Res*. 2000, **28**, 27-30.

[31] MetaCyc: a multiorganism database of metabolic pathways and enzymes. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD. *Nucleic Acids Res*. 2004, **32**, D438-42.

[32] High-throughput generation, optimization and analysis of genome-scale metabolic models. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. *Nat Biotechnol*. 2010, **28**, 977-82.

[33] Towards multidimensional genome annotation. Reed JL, Famili I, Thiele I, Palsson BO. *Nat Rev Genet*. 2006, **7**, 130-41.

[34] The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Feist AM, Palsson BØ. *Nat Biotechnol*. 2008, **26**, 659-67.

[35] Can a biologist fix a radio?--Or, what I learned while studying apoptosis. Lazebnik Y. *Cancer Cell*. 2002, **2**, 179-82.

[36] Control of metabolic systems. Burns JA, Cornish-Bowden A, Groen AK, Heinrich R, Kacser H, Porteous JW, Rapoport SM, Rapoport TA, Stucki JW, Tager JM, Wanders RJA, Westerhoff HV. *Trends Biochem. Sci*. 1985, **10**, 16.

[37] GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. Mendes P. *Comput Appl Biosci*. 1993, **9**, 563-71.

[38] What is flux balance analysis? Orth JD, Thiele I, Palsson BØ. *Nat Biotechnol*. 2010, **28**, 245-8.

[39] Enzyme kinetics and computational modeling for systems biology. Mendes P, Messiha H, Malys N, Hoops S. *Methods Enzymol*. 2009, **467**, 583-99.

[40] Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL. *Eur J Biochem*. 2000, **267**, 5313-29.

[41] BRENDA, enzyme data and metabolic information. Schomburg I, Chang A, Schomburg D. *Nucleic Acids Res*. 2002, **30**, 47-9.

[42] Storing and annotating of kinetic data. Rojas I, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Wittig U. *In Silico Biol*. 2007, **7**, S37-44.

[43] Global analysis of protein expression in yeast. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. *Nature*. 2003, **425**, 737-741.

[44] Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. *Cell*. 2009, **138**, 795-806.

[45] Comprehensive quantitative analysis of central carbon and amino-acid metabolism in Saccharomyces cerevisiae under multiple conditions by targeted proteomics. Costenoble R, Picotti P, Reiter L, Stallmach R, Heinemann M, Sauer U, Aebersold R. *Mol Syst Biol*. 2011, **7**, 464.

[46] Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. Pratt JM, Simpson DM, Doherty MK, Rivers J, Gaskell SJ, Beynon RJ. *Nat Protoc*. 2006, **1**, 1029-43.

[47] Reconstruction annotation jamborees: a community approach to systems biology. Thiele I, Palsson BØ. *Mol Syst Biol*. 2010, **6**, 361.

[48] Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Duarte NC, Herrgård MJ, Palsson BØ. *Genome Res*. 2004, **14**, 1298-309.

[49] Detection of stoichiometric inconsistencies in biomolecular models. Gevorgyan A, Poolman MG, Fell DA. *Bioinformatics*. 2008, **24**, 2245-51.

[50] The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Edwards JS, Palsson BO. *Proc Natl Acad Sci U S A*. 2000, **97**, 5528–33.

[51] Simple constrained-optimization view of acetate overflow in E. coli. Majewski RA, Domach MM. *Biotechnol Bioeng*. 1990, **35**, 732-8.

[52] A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. *Mol Syst Biol*. 2007, **3**, 121.

[53] CellML: its future, present and past. Lloyd CM, Halstead MD, Nielsen PF. *Prog Biophys Mol Biol*. 2004, **85**, 433-50.

[54] The BioPAX community standard for pathway data sharing. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S,

Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. *Nat Biotechnol*. 2010, **28**, 935-42.

[55] The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. 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. *Bioinformatics*. 2003, **19**, 524-31.

[56] Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgård MJ. *Nat Protoc*. 2007, **2**, 727-38.

[57] BioMet Toolbox: genome-wide analysis of metabolism. Cvijovic M, Olivares-Hernández R, Agren R, Dahr N, Vongsangnak W, Nookaew I, Patil KR, Nielsen J. *Nucleic Acids Res*. 2010, **38**, W144-9.

[58] FASIMU: flexible software for flux-balance computation series in large metabolic networks. Hoppe A, Hoffmann S, Gerasch A, Gille C, Holzhütter HG. *BMC Bioinformatics*. 2011, **12**, 28.

[59] Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Förster J, Famili I, Fu P, Palsson BØ, Nielsen *J. Genome Res*. 2003, **13**, 244-53.

[60] Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Kuepfer L, Sauer U, Blank LM. *Genome Res*. 2005, **15**, 1421-30.

[61] Towards whole-body systems physiology. Kuepfer L. *Mol Syst Biol*. 2010, **6**, 409.

[62] Minimum information requested in the annotation of biochemical models (MIRIAM). Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H,

Shapiro B, Snoep JL, Spence HD, Wanner BL. *Nat Biotechnol*. 2005, **23**, 1509-15.

[63] ChEBI: a database and ontology for chemical entities of biological interest. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. *Nucleic Acids Res*. 2008, **36**, D344-50.

[64] Chemical Entities of Biological Interest: an update. de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C. *Nucleic Acids Res*. 2010, **38**, D249-54.

[65] PubMed: bridging the information gap. McEntyre J, Lipman D. *CMAJ*. 2001, **164**, 1317-9.

[66] Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. *Nat Genet*. 2000, **25**, 25-9.

[67] PubChem: integrated platform of small molecules and biological activities. Bolton EE, Wang Y, Thiessen PA, Bryant SH. *Annu. Rep. Comput. Chem*. 2008, **4**, 217–41.

[68] HMDB: the Human Metabolome Database. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. *Nucleic Acids Res*. 2007, **35**, D521-6.

[69] SGD: Saccharomyces Genome Database. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. *Nucleic Acids Res*. 1998, **26**, 73-9.

[70] UniProt: the Universal Protein knowledgebase. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R,

Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. *Nucleic Acids Res*. 2004, **32**, D115-9.

[71] Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Webb EC. *San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press*. 1992. ISBN 0-12-227164-5.

[72] Model storage, exchange and integration. Le Novère N. *BMC Neurosci*. 2006, **7**, S11.

[73] A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB. *Nat Biotechnol*. 2008, **26**, 1155-60.

[74] A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RM, Hsiung CA, De Keersmaecker SC, Liao YC, Marchal K, Mo ML, Özdemir E, Raghunathan A, Reed JL, Shin SI, Sigurbjörnsdóttir S, Steinmann J, Sudarsan S, Swainston N, Thijs IM, Zengler K, Palsson BO, Adkins JN, Bumann D. *BMC Syst Biol*. 2011, **5**, 8.

[75] Nutrient control of eukaryote cell growth: a systems biology study in yeast. Gutteridge A, Pir P, Castrillo JI, Charles PD, Lilley KS, Oliver SG. *BMC Biol*. 2010, **8**, 68.

[76] Social engineering for virtual 'big science' in systems biology. Kitano H, Ghosh S, Matsuoka Y. *Nat Chem Biol*. 2011, **7**, 323-6.

[77] Reconstruction and validation of RefRec: a global model for the yeast molecular interaction network. Aho T, Almusa H, Matilainen J, Larjo A,

Ruusuvuori P, Aho KL, Wilhelm T, Lähdesmäki H, Beyer A, Harju M, Chowdhury S, Leinonen K, Roos C, Yli-Harja O. *PLoS One*. 2010, **5**, e10662.

[78] Customizable views on semantically integrated networks for systems biology. Weile J, Pocock M, Cockell SJ, Lord P, Dewar JM, Holstein EM, Wilkinson D, Lydall D, Hallinan J, Wipat A. *Bioinformatics*. 2011, **27**, 1299-306.

[79] Towards genome-scale signalling network reconstructions. Hyduke DR, Palsson BØ. *Nat Rev Genet*. 2010, **11**, 297-307.

[80] The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, Nielsen J, Bhumiratana S. *BMC Syst Biol*. 2008, **2**, 71.

[81] LIPID MAPS online tools for lipid research. Fahy E, Sud M, Cotter D, Subramaniam S. *Nucleic Acids Res*. 2007, **35**, W606-12.

[82] The Ensembl genome database project. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. *Nucleic Acids Res*. 2002, **30**, 38-41.

[83] Functional profiling of the Saccharomyces cerevisiae genome. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke

JD, Snyder M, Philippsen P, Davis RW, Johnston M. *Nature*. 2002, **418**, 387-91.

[84] Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segrè D. *Genome Biol*. 2008, **9**, R140.

[85] Further developments towards a genome-scale metabolic model of yeast. Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD, Mendes P. *BMC Syst Biol*. 2010, **4**, 145.

[86] Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. Zomorrodi AR, Maranas CD. *BMC Syst Biol*. 2010, **4**, 178.

[87] Optimization based automated curation of metabolic reconstructions. Satish Kumar V, Dasika MS, Maranas CD. *BMC Bioinformatics*. 2007, **8**, 212.

[88] GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. Kumar VS, Maranas CD. *PLoS Comput Biol*. 2009, **5**, e1000308.

[89] A protocol for generating a high-quality genome-scale metabolic reconstruction. Thiele I, Palsson BØ. *Nat Protoc*. 2010, **5**, 93-121.

[90] metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Pinney JW, Shirley MW, McConkey GA, Westhead DR. *Nucleic Acids Res*. 2005, **33**, 1399-409.

[91] The Pathway Tools software. Karp PD, Paley S, Romero P. *Bioinformatics*. 2002, **18**, S225-32.

[92] KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. Moutselos K, Kanaris I, Chatziioannou A, Maglogiannis I, Kolisis FN. *BMC Bioinformatics*. 2009, **10**, 324.

[93] MetNetMaker: a free and open-source tool for the creation of novel metabolic networks in SBML format. Forth T, McConkey GA, Westhead DR. *Bioinformatics*. 2010, **26**, 2352-3.

[94] Challenges to be faced in the reconstruction of metabolic networks from public databases. Poolman MG, Bonde BK, Gevorgyan A, Patel HH, Fell DA. *Syst Biol (Stevenage)*. 2006, **153**, 379-84.

[95] LibSBML: an API library for SBML. Bornstein BJ, Keating SM, Jouraku A, Hucka M. *Bioinformatics*. 2008, **24**, 880-1.

[96] MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. Laibe C, Le Novère N. *BMC Syst Biol*. 2007, **1**, 58.

[97] Database resources of the National Center for Biotechnology Information. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. *Nucleic Acids Res*. 2009, **37**, D5-15.

[98] libAnnotationSBML: a library for exploiting SBML annotations. Swainston N, Mendes P. *Bioinformatics*. 2009, **25**, 2292-3.

[99] Saint: a lightweight integration environment for model annotation. Lister AL, Pocock M, Taschuk M, Wipat A. *Bioinformatics*. 2009, **25**, 3026-7.

[100] STRING 8--a global view on proteins and their functional interactions in 630 organisms. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. *Nucleic Acids Res*. 2009, **37**, D412-6.

[101] Annotation and merging of SBML models with semanticSBML. Krause F, Uhlendorf J, Lubitz T, Schulz M, Klipp E, Liebermeister W. *Bioinformatics*. 2010, **26**, 421-2.

[102] BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. Li C, Donizelli M, Rodriguez N, Dharuri

H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, Snoep JL, Hucka M, Le Novère N, Laibe C. *BMC Syst Biol*. 2010, **4**, 92.

[103] The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. Swainston N, Smallbone K, Mendes P, Kell DB, Paton NW. *J Integr Bioinform*. 2011, **8**, 186.

[104] SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Weininger D. *J. Chem. Inf. Comput. Sci*. 1998, **28**, 31-36.

[105] SBML Level 3 Package Proposal: Annotation. Waltemath D, Swainston N, Lister A, Bergmann F, Henkel R, Hoops S, Hucka M, Juty N, Keating S, Knuepfer C, Krause F, Laibe C, Liebermeister W, Lloyd C, Misirli G, Schulz M, Taschuk M, Le Novère N. *Nature Precedings*, 2011 http://dx.doi.org/10.1038/npre.2011.5610.1.

[106] Towards building the silicon cell: a modular approach. Snoep JL, Bruggeman F, Olivier BG, Westerhoff HV. *Biosystems*. 2006, **83**, 207-16.

[107] Die Kinetik der Invertinwirkung. Michaelis L, Menten ML. *Biochem. Z*. 1913, **49**, 333-369.

[108] qPIPSA: relating enzymatic kinetic parameters and interaction fields. Gabdoulline RR, Stein M, Wade RC. *BMC Bioinformatics*. 2007, **8**, 373.

[109] SYCAMORE--a systems biology computational analysis and modelling research environment. Weidemann A, Richter S, Stein M, Sahle S, Gauges R, Gabdoulline R, Surovtsova I, Semmelrock N, Besson B, Rojas I, Wade R, Kummer U. *Bioinformatics*. 2008, **24**, 1463-4.

[110] Measuring enzyme activities under standardized in vivo-like conditions for systems biology. van Eunen K, Bouwman J, Daran-Lapujade P, Postmus J, Canelas AB, Mensonides FI, Orij R, Tuzun I, van den Brink J, Smits GJ, van Gulik WM, Brul S, Heijnen JJ, de Winde JH, de Mattos MJ, Kettner C, Nielsen J, Westerhoff HV, Bakker BM. *FEBS J*. 2010, **277**, 749-60.

[111] COPASI--a COmplex PAthway SImulator. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U. *Bioinformatics*. 2006, **22**, 3067-74.

[112] Uniform reporting standards for enzyme activity data. Kettner C. *FEBS J*. 2009, **276**, Supplement s1, 108-109.

[113] Integration of CellDesigner and SABIO-RK. Funahashi A, Jouraku A, Matsuoka Y, Kitano H. *In Silico Biol*. 2007, **7**, S81-90.

[114] Enzyme kinetics informatics: from instrument to browser. Swainston N, Golebiewski M, Messiha HL, Malys N, Kania R, Kengne S, Krebs O, Mir S, Sauer-Danzwith H, Smallbone K, Weidemann A, Wittig U, Kell DB, Mendes P, Müller W, Paton NW, Rojas I. *FEBS J*. 2010, **277**, 3769–79.

[115] Understanding the languages of cells - network modelling in metabolic systems biology and biotechnology: Why, how and whither. Kell DB. *Biochemist*. 2011, **33**, 4-7.

[116] Systematic integration of experimental data and models in systems biology. Li P, Dada JO, Jameson D, Spasic I, Swainston N, Carroll K, Dunn W, Khan F, Malys N, Messiha HL, Simeonidis E, Weichart D, Winder C, Wishart J, Broomhead DS, Goble CA, Gaskell SJ, Kell DB, Westerhoff HV, Mendes P, Paton NW. *BMC Bioinformatics*. 2010, **11**, 582.

[117] Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Mendes P, Kell D. *Bioinformatics*. 1998, **14**, 869-83.

[118] Mass spectrometry-based proteomics turns quantitative. Ong SE, Mann M. *Nat Chem Biol*. 2005, **1**, 252-62.

[119] Rapid identification of proteins by peptide-mass fingerprinting. Pappin DJ, Hojrup P, Bleasby AJ. *Curr. Biol*. 1993, **3**, 327–32.

[120] Mass spectrometry-based proteomics. Aebersold R, Mann M. *Nature*. 2003, **422**, 198-207.

[121] Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. *Mol Cell Proteomics*. 2002, **1**, 376-86.

[122] Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Ross PL, Huang YN, Marchese JN,

Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. *Mol Cell Proteomics*. 2004, **3**, 1154-69.

[123] Growth control of the eukaryote cell: a systems biology study in yeast. Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst D, O'Donoghue K, Hester SS, Dunkley TP, Hart SR, Swainston N, Li P, Gaskell SJ, Paton NW, Lilley KS, Kell DB, Oliver SG. *J Biol*. 2007, **6**, 4.

[124] PRIDE: the proteomics identifications database. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. *Proteomics*. 2005, **5**, 3537-45.

[125] Democratizing proteomics data. Anon. *Nat Biotechnol*. 2007, **25**, 262.

[126] The PRIDE proteomics identifications database: data submission, query, and dataset comparison. Jones P, Côté R. *Methods Mol Biol*. 2008, **484**, 287-303.

[127] Probability-based protein identification by searching sequence databases using mass spectrometry data. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis*. 1999, **20**, 3551-67.

[128] The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. Côté RG, Jones P, Apweiler R, Hermjakob H. *BMC Bioinformatics*. 2006, **7**, 97.

[129] i-Tracker: for quantitative proteomics using iTRAQ. Shadforth IP, Dunkley TP, Lilley KS, Bessant C. *BMC Genomics*. 2005, **6**, 145.

[130] Implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. Orchard S, Albar JP, Deutsch EW, Eisenacher M, Binz PA, Hermjakob H. *Proteomics*. 2010, **10**, 1895-8.

[131] An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ. Siepen JA, Swainston N, Jones AR, Hart SR, Hermjakob H, Jones P, Hubbard SJ. *Proteome Sci*. 2007, **5**, 4.

[132] Recent developments in public proteomic MS repositories and pipelines. Mead JA, Bianco L, Bessant C. *Proteomics*. 2009, **9**, 861-81.

[133] A guide to the Proteomics Identifications Database proteomics data repository. Vizcaíno JA, Côté R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L. *Proteomics*. 2009, **9**, 4276-83.

[134] Differential protein expression analysis using stable isotope labeling and PQD linear ion trap MS technology. Armenta JM, Hoeschele I, Lazar IM. *J Am Soc Mass Spectrom*. 2009, **20**, 1287-302.

[135] iQuantitator: a tool for protein expression inference using iTRAQ. Schwacke JH, Hill EG, Krug EL, Comte-Walters S, Schey KL. *BMC Bioinformatics*. 2009, **10**, 342.

[136] PRIDE: new developments and new datasets. Jones P, Côté RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D, Hermjakob H. *Nucleic Acids Res*. 2008, **36**, D878-83.

[137] Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy. Brownridge P, Holman SW, Gaskell SJ, Grant CM, Harman VM, Hubbard SJ, Lanthaler K, Lawless C, O'cualain R, Sims P, Watkins R, Beynon RJ. *Proteomics*. 2011, **11**, 2957-70.

[138] Computational prediction of proteotypic peptides for quantitative proteomics. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. *Nat Biotechnol*. 2007, **25**, 125-31.

[139] MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. Mead JA, Bianco L, Ottone V, Barton C, Kay RG, Lilley KS, Bond NJ, Bessant C. *Mol Cell Proteomics*. 2009, **8**, 696-705.

[140] PRIDE Converter: making proteomics data-sharing easy. Barsnes H, Vizcaíno JA, Eidhammer I, Martens L. *Nat Biotechnol*. 2009, **27**, 598-9.

[141] SILACAnalyzer - A Tool for Differential Quantitation of Stable Isotope Derived Data. Lars Nilse L, Sturm M, Trudgian D, Salek M, Sims PFG, Carroll KM, Hubbard SJ. *Computational Intelligence Methods for Bioinformatics And Biostatistics. Lecture Notes in Computer Science*. 2010, **6160**, 45-55.

[142] mzML: a single, unifying data format for mass spectrometer output. Deutsch E. *Proteomics*. 2008, **8**, 2776-7.

[143] A QconCAT informatics pipeline for the analysis, visualisation and sharing of absolute quantitative proteomics data. Swainston N, Jameson D, Carroll K. *Proteomics*. 2011, **11**, 329–33.

[144] Derivatization or not: a choice in quantitative proteomics. Yao X. *Anal Chem*. 2011, **83**, 4427-39.

[145] Integrative Information Management for Systems Biology. Swainston N, Jameson D, Li P, Spasić I, Mendes P, Paton NW. *In proceedings of the 7th International workshop on Data Integration in the Life Sciences 2010 (DILS'10), Gothenburg, Sweden. Lecture Notes in Computer Science*. 2010, **6254**, 164-78.

[146] Workflows for Information Integration in the Life Sciences. Missier P, Paton NW, Li P. *Search Computing: Trends and Developments, S. Ceri and M. Brambilla (eds), Springer*. 2011, 215-26.

[147] Design and Architecture of Web Services for Simulation of Biochemical Systems. Dada JO, Mendes P. *In proceedings of the 6th International workshop on Data Integration in the Life Sciences 2009 (DILS'09), Manchester, UK. Lecture Notes in Computer Science*. 2009, **5647**, 182-95.

[148] Taverna: a tool for the composition and enactment of bioinformatics workflows. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. *Bioinformatics*. 2004, **20**, 3045-54.

[149] Software that goes with the flow in systems biology. Hucka M, Le Novère N. *BMC Biol*. 2010, **8**, 140.

[150] MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. Spasić I, Dunn WB, Velarde G, Tseng A, Jenkins H, Hardy N, Oliver SG, Kell DB. *BMC Bioinformatics*. 2006, **7**, 281.

[151] Data integration for dynamic and sustainable systems biology resources: challenges and lessons learned. Sullivan DE, Gabbard JL Jr, Shukla M, Sobral B. *Chem Biodivers*. 2010, **7**, 1124-41.

[152] A roadmap for the establishment of standard data exchange structures for metabolomics. Hardy NW, Taylor CF. *Metabolomics*. 2007, **3**, 243-8.

[153] ArrayExpress--a public repository for microarray gene expression data at the EBI. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. *Nucleic Acids Res*. 2003, **31**, 68-71.

[154] Computational strategies for metabolite identification in metabolomics. Wishart DS. *Bioanalysis*. 2009, **1**, 1579-96.

[155] An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Eng JK, McCormack AL, Yates JR 3rd. J. Am*. Soc. Mass Spectrom*. 1994, **5**, 976–89.

[156] A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Fenyö D, Beavis RC. *Anal Chem*. 2003, **75**, 768-74.

[157] 13C metabolic flux analysis. Wiechert W. *Metab Eng*. 2001, **3**, 195-206.

[158] Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum. Wiback SJ, Mahadevan R, Palsson BØ. *Biotechnol Bioeng*. 2004, **86**, 317-31.

[159] FiatFlux--a software for metabolic flux analysis from 13C-glucose experiments. Zamboni N, Fischer E, Sauer U. *BMC Bioinformatics*. 2005, **6**, 209.

[160] Intuitive Visualization and Analysis of Multi-Omics Data and Application to Escherichia coli Carbon Metabolism. Enjalbert B, Jourdan F, Portais JC. *PLoS One*. 2011, **6**, e21318.

[161] iMAT: an integrative metabolic analysis tool. Zur H, Ruppin E, Shlomi T. *Bioinformatics*. 2010, **26**, 3140-2.

[162] Network-based prediction of human tissue-specific metabolism. Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. *Nat Biotechnol*. 2008, **26**, 1003-10.

[163] Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study. Daran-Lapujade P, Jansen ML, Daran JM, van Gulik W, de Winde JH, Pronk JT. *J Biol Chem*. 2004, **279**, 9125-38.

[164] Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Famili I, Forster J, Nielsen J, Palsson BO. *Proc Natl Acad Sci U S A*. 2003, **100**, 13134-9.

[165] Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Jerby L, Shlomi T, Ruppin E. *Mol Syst Biol*. 2010, **6**, 401.

[166] Global reconstruction of the human metabolic network based on genomic and bibliomic data. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. *Proc Natl Acad Sci U S A*. 2007, **104**, 1777-82.

[167] Web-based kinetic modelling using JWS Online. Olivier BG, Snoep JL. *Bioinformatics*. 2004, **20**, 2143-4.

[168] MEMOSys: Bioinformatics platform for genome-scale metabolic models. Pabinger S, Rader R, Agren R, Nielsen J, Trajanoski Z. *BMC Syst Biol*. 2011, **5**, 20.

[169] rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. Thorleifsson SG, Thiele I. *Bioinformatics*. 2011, **27**, 2009-10.

[170] The Systems Biology Graphical Notation. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H. *Nat Biotechnol*. 2009, **27**, 735-41.

[171] CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Funahashi A, Tanimura N, Morohashi M, Kitano H. *BIOSILICO*. 2003, **1**, 159-62.

[172] Payao: a community platform for SBML pathway model curation. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H. *Bioinformatics*. 2010, **26**, 1381-3.

[173] Arcadia: a visualization tool for metabolic pathways. Villéger AC, Pettifer SR, Kell DB. *Bioinformatics*. 2010, **26**, 1470-1.

[174] SED-ML --- An XML Format for the Implementation of the MIASE Guidelines. Köhn D, Le Novère N. *Lect Notes Comput Sci.* 2008, **5307**, 176–90.

[175] Minimum Information About a Simulation Experiment (MIASE). Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, Chelliah V, Cooling MT, Cooper J, Crampin EJ, Garny A, Hoops S, Hucka M, Hunter P, Klipp E, Laibe C, Miller AK, Moraru I, Nickerson D, Nielsen P, Nikolski M, Sahle S, Sauro HM, Schmidt H, Snoep JL, Tolle D, Wolkenhauer O, Le Novère N. *PLoS Comput Biol*. 2011, **7**, e1001122.

[176] SBRML: a markup language for associating systems biology data with models. Dada JO, Spasić I, Paton NW, Mendes P. *Bioinformatics*. 2010, **26**, 932-8.

[177] Comparing simulation results of SBML capable simulators. Bergmann FT, Sauro HM. *Bioinformatics*. 2008, **24**, 1963-5.

[178] Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H, Doyle J, Kitano H. *OMICS*. 2003, **7**, 355-72.

Publications

Publication 1

A consensus yeast metabolic network obtained from a community approach to systems biology.

Herrgård MJ*, **Swainston N***, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB.

Nat. Biotechnol. 2008, **26**, 1155-60.

**Equal contribution.*

A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology

Markus J Herrgård[1,19,20], Neil Swainston[2,3,20], Paul Dobson[3,4], Warwick B Dunn[3,4], K Yalçin Arga[5], Mikko Arvas[6], Nils Blüthgen[3,7], Simon Borger[8], Roeland Costenoble[9], Matthias Heinemann[9], Michael Hucka[10], Nicolas Le Novère[11], Peter Li[2,3], Wolfram Liebermeister[8], Monica L Mo[1], Ana Paula Oliveira[12], Dina Petranovic[12,19], Stephen Pettifer[2,3], Evangelos Simeonidis[3,7], Kieran Smallbone[3,13], Irena Spasić[2,3], Dieter Weichart[3,4], Roger Brent[14], David S Broomhead[3,13], Hans V Westerhoff[3,7,15], Betül Kırdar[5], Merja Penttilä[6], Edda Klipp[8], Bernhard Ø Palsson[1], Uwe Sauer[9], Stephen G Oliver[3,16], Pedro Mendes[2,3,17], Jens Nielsen[12,18] & Douglas B Kell*[3,4]

Genomic data allow the large-scale manual or semi-automated assembly of metabolic network reconstructions, which provide highly curated organism-specific knowledge bases. Although several genome-scale network reconstructions describe *Saccharomyces cerevisiae* metabolism, they differ in scope and content, and use different terminologies to describe the same chemical entities. This makes comparisons between them difficult and underscores the desirability of a consolidated metabolic network that collects and formalizes the 'community knowledge' of yeast metabolism. We describe how we have produced a consensus metabolic network reconstruction for *S. cerevisiae*. In drafting it, we placed special emphasis on referencing molecules to persistent databases or using database-independent forms, such as SMILES or InChI strings, as this permits their chemical structure to be represented unambiguously and in a manner that permits automated reasoning. The reconstruction is readily available via a publicly accessible database and in the Systems Biology Markup Language (http://www.comp-sys-bio.org/yeastnet). It can be maintained as a resource that serves as a common denominator for studying the systems biology of yeast. Similar strategies should benefit communities studying genome-scale metabolic networks of other organisms.

Accurate representation of biochemical, metabolic and signaling networks by mathematical models is a central goal of integrative systems biology. This undertaking can be divided into four stages[1]. The first is a qualitative stage in which are listed all the reactions that are known to occur in the system or organism of interest; in the modern era, and especially for metabolic networks, these reaction lists are often derived in part from genomic annotations[2,3] with curation based on literature ('bibliomic') data[4]. A second stage, again qualitative, adds known effectors, whereas the third and fourth stages—essentially amounting to molecular enzymology—include the known kinetic rate equations and the values of their parameters. Armed with such information, it is then possible to provide a stochastic or ordinary differential equation model of the entire metabolic network of interest. An attractive feature of metabolism, for the purposes of modeling, is that, in contrast to signaling pathways, metabolism is subject to direct thermodynamic and (in particular) stoichiometric constraints[3]. Our focus here is on the first two stages of the reconstruction process, especially as it pertains to the mapping of experimental metabolomics data onto metabolic network reconstructions.

Besides being an industrial workhorse for a variety of biotechnological products, *S. cerevisiae* is a highly developed model organism for biochemical, genetic, pharmacological and post-genomic studies[5]. It is especially attractive because of the availability of its genome sequence[6], a whole series of bar-coded deletion[7,8] and other[9] strains, extensive experimental 'omics data[10–14] and the ability to grow it for extended periods under highly controlled conditions[15]. The very active scientific community that works on *S. cerevisiae* has a history of collaborative research projects that have led to substantial advances in our understanding of eukaryotic biology[6,8,13,16,17]. Furthermore, yeast metabolic physiology has been the subject of intensive study and most of the components of the yeast metabolic network are relatively well characterized. Taken together, these factors make yeast metabolism an attractive topic to test a community approach to build models for systems biology.

Several groups[18–21] have reconstructed the metabolic network of yeast from genomic and literature data and made the reconstructions freely available. However, due to different approaches used to create them, as well as different interpretations of the literature, the existing reconstructions have many differences. Additionally, the naming of metabolites and enzymes in the existing reconstructions was, at best, inconsistent, and there were no systematic annotations of the chemical species in the form of links to external databases that store chemical compound information. This lack of model annotation complicated the use of the models for data analysis and integration. Members of the yeast systems biology community therefore recognized that a single 'consensus' reconstruction and annotation of the metabolic network was highly desirable as a starting point for further investigations.

A crucial factor that enabled the building of a consensus network reconstruction is the ability to describe and exchange biochemical network

*A list of affiliations appears at the end of the paper.

Published online 9 October 2008; doi:10.1038/nbt1492

models in a standard format, the Systems Biology Markup Language (SBML; http://www.sbml.org/)[22]. The SBML format is employed by most commonly used software applications for visualizing, simulating and analyzing biochemical networks, and also in pathway databases. SBML also provides the necessary standardized means ('Minimum Information Requested in the Annotation of biochemical Models' or MIRIAM[23]) to annotate models with information that is required to identify network components uniquely, including metabolites, proteins and genes. Representing the consensus metabolic network reconstruction in a MIRIAM-compatible SBML format allows widespread use of the reconstruction and assists in its continued curation, expansion and revision.

We developed this consensus reconstruction using a 'jamboree' approach—a large, focused work meeting, where we defined the protocol for the curation process as well as resolving the majority of discrepancies between the existing reconstructions. The jamboree event was followed by an extended process of curation of remaining discrepancies and careful annotation of components of the reconstructions by a smaller group of people. The overall goal of the effort was, by careful curation and comprehensive annotation of the network model and its components, to make the consensus reconstruction useful for the broadest possible set of users. The general reconstruction could then be used directly in bioinformatics applications aimed at integration of, for example, metabolomics and proteomics data or as a starting point for building predictive models using a number of different approaches[24,25], and for other purposes outlined below.

Here we describe how an initial 'community consensus' reconstruction of the yeast metabolic network was carried out. We make some further proposals for how this reconstruction of the yeast metabolic network may evolve as more information is acquired. We also discuss the possibility of using a similar approach to build consensus models of metabolic and other networks in other organisms.

Consensus reconstruction

As a starting point for the development of a consensus reconstruction, we chose two separately developed freely available metabolic network reconstructions, iMM904 (see http://www.cmb.dtu.dk/Forskning/Software/models.aspx and http://gcrg.ucsd.edu/In_Silico_Organisms/Yeast) and iLL672 (ref. 20), containing 904 and 672 yeast genes, respectively. We have also placed relevant files in SBML format on the website http://www.comp-sys-bio.org/yeastnet. Both of these reconstructions were derived from the first genome-scale metabolic network reconstruction for yeast iFF708 (ref. 18; for the basis of this terminology see ref. 26), but the process of curating the original reconstruction was substantially different for the two derived reconstructions. The iMM904 reconstruction has eight different compartments and was developed by curating and expanding an earlier reconstruction, iND750 (ref. 19). In contrast, the iLL672 reconstruction[20] was directly derived from iFF708 by extensively curating the reconstruction to improve the ability of the flux balance model derived from the reconstruction to predict gene deletion phenotypes[27]. It should be noted that yeast metabolic pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[28] and the Saccharomyces Genome Database (SGD)[29,30] databases were used to establish the starting point for building the original iFF708 reconstruction and also for curating the iLL672 and iMM904 reconstructions. Hence, the information from early versions of these two reconstructions is included implicitly in the consensus reconstruction.

Table 1 Comparison of starting-point reconstructions

	iMM904	iLL672	Common	iMM904 only	iLL672 only
Metabolites	713	452	444	269	8
Reactions	1,402	743	566[a]	836	177
Genes	904	659	646	258	13
Compartments	8	2	2	6	0

[a]Reaction comparisons were done by considering every reaction to be reversible and without taking into account water and extracellular or intracellular protons (explicitly accounted for in iMM904).

Due to the lack of common metabolite names and annotations, the comparison of the two starting-point reconstructions required first manually defining the correspondences between metabolites. After these had been assigned, the overall metabolite and reaction content of the two reconstructions could be compared (**Table 1**). The majority of metabolites (444) were found in both reconstructions, whereas 8 were found only in iLL672 and 269 only in iMM904. In terms of reactions, 566 were in both reconstructions, 177 were only in iLL672 and 836 only in iMM904. The large number of additional reactions in iMM904 is mostly due to the expanded number of compartments represented in this reconstruction.

The jamboree was held at The University of Manchester, UK, in April 2007. The comparison between the iLL672 and iMM904 reconstructions, proposed at a meeting of the Yeast Systems Biology Network (http://www.ysbn.eu/) in Helsinki, Finland, in June 2006, formed the starting point for the reconstruction (**Table 1**). The three-day event in Manchester concentrated on three separate areas: (i) defining standards for curation as well as for representation of the annotated reconstruction in SBML, (ii) annotating the metabolites with reference to external compound databases and (iii) resolving discrepancies between the reaction-metabolite sets in the two reconstructions. The presence of experts in fields such as yeast genetics and physiology, systems modeling, metabolomics, standards (SBML/MIRIAM/metabolomics), and database or ontology development allowed the group to make good progress in all three areas. The annotation and curation was aided by a version of the B-Net database[31], and is provided in SBML form (**Supplementary Table 1** online). After the jamboree, a subgroup of the authors verified the curation and annotation, and resolved the remaining discrepancies between models. Below, we discuss some of the major components of the curation and annotation processes.

Metabolite-naming conventions

The initial comparison made it very clear that the naming conventions used in the two models were completely different, such that it was difficult in some cases even for experts to know which chemical entities were being referred to. Moreover, some of the reactions involved 'generic' structures (molecules with R-groups or so-called 'Markush' structures), which are not effectively represented in stoichiometric metabolic models, while certain named entities represented 'composite' substances such as mixtures of different lipids or 'biomass'. Without standardized names, it is extremely hard to enable computer software to reason about the similarities and differences between different models[32–37]. This is even more problematic in the case of reconstructions of the larger human metabolic network[4,38].

However, as SBML allows one to annotate species such as metabolites with external references, we related them to molecules in the 'chemical entities of biological interest' (ChEBI)[39], KEGG[28] and PubChem[40] databases, and identified them precisely using database-independent representations of small molecules, such as 'simplified molecular input line entry system' (SMILES)[41] and international chemical identifier (InChI)[36,42] representations. We took advantage of this aspect of SBML to identify and annotate manually which chemical species were being described. In general, we searched these databases with the contents of the species' name attribute field in the SBML representation or by the chemical formula of the compound sought. The order of annotation was such that we annotated metabolite species using ChEBI identifiers and InChI strings, where possible. If these did not exist or could not be resolved, we used KEGG IDs— or, in two cases, Human Metabolome Database (HMDB)

identifiers[43]—followed by PubChem IDs and finally PubMed references. This generated, for the first time, a representation that allows computational comparisons to be performed.

Because some individual molecules have multiple states (e.g., because of acid-base reactions), it would be desirable to use the chemical entities believed to be most common at the pH of the relevant compartment. However, in this version of the consensus reconstruction, all species are assumed to be in the form that corresponds to the most common protonation state at pH 7.2. Whenever possible, the metabolites were annotated with a database entry with the correct protonation state. However, in several cases, the databases only contained the metabolite in a neutral form or otherwise in an incorrect or incorrectly annotated protonation state.

Annotation of large-scale metabolic models in SBML

Although large-scale metabolic network reconstructions and models are now commonly represented in SBML, there has not thus far been a standard way to annotate these models. As part of the consensus reconstruction effort, we tried to develop such a standard that is compliant with MIRIAM[23]. Whereas the annotation of metabolites is quite straightforward, standardized annotation of the reaction content (molecules and reactions) of the reconstructed network proved to be more involved.

Where possible, we annotated reactions using literature references encoded as PubMed IDs, using the MIRIAM- and SBML-compliant "isDescribedBy" 'resource description framework' (RDF; see http://www.w3.org/TR/REC-rdf-syntax/) annotation tag. In addition, reaction annotations include modifiers (enzymes/enzyme complexes) where possible. If a given reaction can be catalyzed by two or more isozymes, we generated an individual reaction for each isozyme (or complex). We represented the formation of protein complexes by separate reactions. Proteins and genes were finally annotated by references to SGD[29] and UniProt[44]. In addition, we annotated cellular compartments using 'Gene Ontology' (GO) terms[45]. In all cases where annotations were used, the MIRIAM[23] web services (http://www.ebi.ac.uk/compneur-srv/miriam-main/mdb?section=ws) were consulted to ensure correct annotation. Examples of fully annotated species and reaction entries are shown in **Figure 1** and in **Supplementary Figure 1** online.

a

```
<species metaid="metaid_M_172" id="M_172" name="ATP" compartment="C_1"
sboTerm="SBO:0000299">
  <annotation>
    <in:inchi xmlns:in="http://biomodels.net/inchi" metaid="M_172_inchi">
InChI=1/C10H16N5O13P3/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6
(16)4(26-10)1-25-30(21,22)28-31(23,24)27-29(18,19)20/h2-4,6-7,10,16-17H,1H2,
(H,21,22)(H,23,24)(H2,11,12,13)(H2,18,19,20)/t4-,6-,7-,10-/m1/s1/f/h18-19,21
,23H,11H2
    </in:inchi>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
      <rdf:Description rdf:about="#metaid_M_172">
        <bqbiol:is>
          <rdf:Bag>
            <rdf:li rdf:resource="urn:miriam:obo.chebi:CHEBI:15422"/>
            <rdf:li rdf:resource="#M_172_inchi"/>
          </rdf:Bag>
        </bqbiol:is>
      </rdf:Description>
    </rdf:RDF>
  </annotation>
</species>
```

b

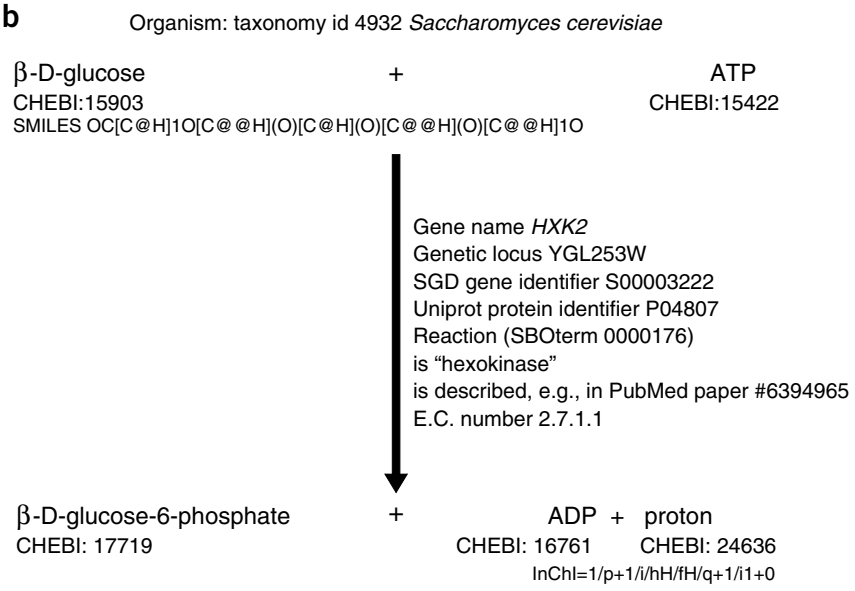

Figure 1 An example of the SBML annotation of a metabolite species using the example of ATP, as used in the reconstruction of the consensus network, illustrating its use of the Systems Biology Ontology (http://www.ebi.ac.uk/sbo/) and its MIRIAM compliance. (**a**) Relevant parts of the SBML code. (**b**) An indication of the kinds of annotations included (for clarity, not all are shown).

Contents of the consensus reconstructions

In all, the resulting consensus network consists of 2,153 species (1,168 metabolites, 832 genes, 888 proteins and 96 catalytic protein complexes) and 1,857 reactions (1,761 metabolic reactions and 96 complex formation reactions). Reactions and species can be localized to 15 compartments (**Table 2**), including membrane compartments. The network contains 664 distinct chemical entities (e.g., ATP present in the nucleus, cytoplasm, Golgi, mitochondrion, peroxisome and vacuole is classified as one chemical species and Mg^{2+} liganding is ignored). Of these distinct chemical entities, 554 are annotated with ChEBI identifiers, 564 with InChI identifiers, 78 with KEGG identifiers, 10 with PubChem identifiers, 2 with HMDB identifiers and only 5 with PubMed references. In addition, 26 compounds are currently not annotated in this way. The majority of these are fatty acyl CoAs or acyl carrier proteins where the corresponding fatty acid is in public databases, but the fatty acyl CoA or acyl carrier protein is currently not deposited (but will be submitted to them).

The network includes 1,312 unique chemical transformations, of which 911 occur within a single compartment and the remaining 401 are transport reactions. The overall distribution of metabolites and reactions between the various compartments in the consensus network is given in **Table 2**. Enzyme Commission (EC) number and PubMed reference annotations are provided for 738 and 478 unique

Table 2 Summary of the consensus reconstruction by cellular compartment

Compartment	Reactions	Metabolites
Cytoplasm	835	590
Extracellular	15	158
Golgi	2	13
Mitochondrion	188	235
Nucleus	30	42
Endoplasmic reticulum (ER)	32	28
Vacuole	2	22
Peroxisome	77	80
Mitochondrial membrane	142	0
Plasma membrane	311	0
Peroxisomal membrane	44	0
ER membrane	17	0
Vacuolar membrane	35	0
Golgi membrane	5	0
Nuclear membrane	26	0

transformations in the network, respectively. Each reaction includes all of its cofactors (sometimes known as 'currency metabolites'), such as ATP, NADH and CoA. In addition, although we recognize that there is a certain arbitrariness about this, we have assigned pathway names for each reaction in the network.

We have removed various reactions from the initial networks, especially where they contained Markush structures or ambiguities. This has led to the underrepresentation of lipids, where there are many combinatorial issues[46]. We anticipate that lipid pathways will be added in the future, but 'lipidomics' experiments will eventually be necessary to define the full complement of lipid species present in *S. cerevisiae*. In a similar vein, composite items such as 'biomass' are excluded. Although these are required for flux balance analysis, our purpose here is to provide the basic inventory of metabolites and network structure that can be used, for example, to compare the network with experimental metabolomics data. This inventory can then form the basis for setting up flux balance models using different assumptions required for setting up these kinds of models, for example, assumptions on the biomass composition, reversibility of reactions and lumping of the reactions into fewer compartments.

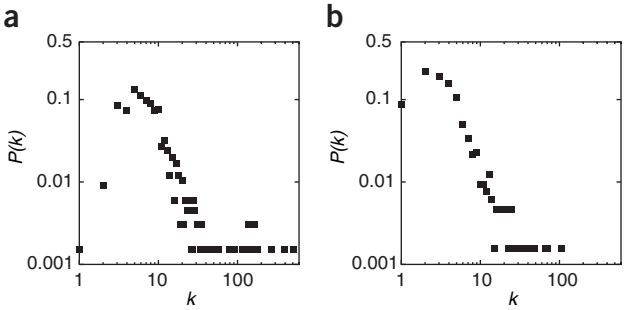

Figure 2 Degree distribution of the metabolic network. The metabolic reaction network was first summarized in a metabolite network, where metabolites are the nodes and one edge links two metabolites that co-occur in a reaction (in any role as substrates or products), as described[48]. For this analysis, transport steps were not considered nor were protein-protein binding reactions. (**a**,**b**) The figures plot the distribution of the degree of connectivity, *P(k)*, expressed as the fraction of metabolites that have *k* links out of the total number of metabolites plotted against the number of links (*k*) in the complete network (**a**) and in a network where the following metabolites were not considered (**b**): {water, proton, carbon dioxide, dioxygen, phosphate3–, diphosphate4–, ammonium, ATP, ADP, AMP, NAD+, NADH, NADP+, NADPH} (to be comparable with the analysis in ref. 48).

Figure 2 depicts the degree distribution[47] of the complete metabolite network, and a version where the currency metabolites were ignored as described earlier[48]. The complete network (**Fig. 2a**) has an average clustering coefficient of 0.742, average node degree of 13.166, characteristic path length of 2.186 and betweenness centralization of 0.3897. The network without currency metabolites (**Fig. 2b**) has an average clustering coefficient of 0.421, average node degree of 5.138, characteristic path length of 4.178 and betweenness centralization of 0.2329. In the full network, the largest value for the shortest distance between any two metabolites ('diameter') is only 4 reaction steps, whereas it is 11 reaction steps (between dTTP and heme A) in the one without 'currency' metabolites. These statistics indicate that the currency metabolites should not be ignored as is sometimes done; without them the network is considerably less connected and several unconnected subnetworks appear, thus leaving some areas of metabolism unconnected from the rest. The center metabolite in the complete network is the proton, whereas in the smaller one it is coenzyme A. **Table 3** lists the top 15 most-connected metabolites of each network.

Dissemination and future curation of the reconstruction

An SBML-encoded version of the base model (with and without compartments) is available at http://www.comp-sys-bio.org/yeastnet. Specifically, the SBML representation of the model is made available under the Creative Commons Attribution-Share Alike 3.0 Unported License (http://www.creativecommons.org/). This is the preferred source for using the complete model with systems biology software. We have tested the SBML using various XML validators, and shown that it loads successfully into the COmplex PAthway SImulator (COPASI)[49] software. COPASI shows that there are 307 mass conservation relations, which were calculated from the stoichiometry matrix using the method of Vallabhajosyula[50], which is now standard in COPASI[49]. We have also loaded the model successfully into some versions of Cytoscape[51] and CellDesigner[52]. The SBML has been checked using libSBML[53,54] (see also http://sbml.org/software/libsbml/).

Recognizing that for many applications only subsets of this model are going to be relevant, we also make it available in an online database that facilitates searching the model. We used the database schema B-Net[31], which already supported all of the features required for our SBML model, including a structured mechanism for MIRIAM annotations. This B-Net representation of the model can be searched using synonyms and it also allows the user to navigate through the network, for example, going from a metabolite to all its reactions, then to the genes that encode the enzymes catalyzing those reactions and so forth. The database is also available at http://www.comp-sys-bio.org/yeastnet.

The B-Net database provides another important function as it is also the preferred means by which the community will be able to edit the model. It will thus be the primary source for the model. As there is no redundancy in the database, any change in any component immediately becomes global. For the time being, editing the model is limited to a few curators to ensure that the current standards are maintained. However, given the major benefits of community annotation[55,56], we have included at the database a mechanism that collects annotations from anyone who wishes to communicate corrections or additions to the model. These annotations will then be reviewed and incorporated into the model for future releases of new versions.

DISCUSSION

We have brought together a large segment of the community engaged in research involving genome-scale metabolic networks of yeast to create a consensus network that is freely available without restrictions and that can form the basis for future improvements. The SBML representation of the reconstruction is freely available under a Creative Commons License, and representations of the network were designed to facilitate future improvements.

Although annotation was semi-automated, a considerable element of manual annotation was still required, especially the parsing of the starting models. One of the biggest problems was the use of nonstandard and often arcane synonyms for referring to the same chemical entity. Several commentators have recognized the difficulties caused by synonyms[4,33,38]. For these purposes, we believe and strongly recommend that the best solution to this synonym problem is to reference chemical entities in persistent databases and with database-independent representations such as SMILES[41] and InChI[42]. Referencing the true chemical entity intended requires detailed consideration of its stereochemistry and the anomeric specificity of reactions in which it is involved, and not all databases have the required level of precision. We also recommend that these networks are first built in an assumption-free manner, and that extra features or assumptions that may be required for specific purposes (e.g., adding composite compounds for flux balance analyses) should only then be introduced and annotated. A further benefit of the jamboree approach is the access to experts necessary to annotate details such as the precise gene-protein relationships underlying specific reactions.

Table 3 Most connected nodes in the metabolite network

Complete metabolite network (as in Fig. 2a)			Abbreviated metabolite network (as in Fig. 2b)		
Metabolite	Degree[a]	Betweenness[b]	Metabolite	Degree[a]	Betweenness[b]
Proton	506	0.391	Coenzyme A	106	0.237
Water	390	0.226	L-glutamate^{1-}	71	0.232
ATP	268	0.099	Acetyl-CoA	66	0.065
Diphosphate^{4-}	181	0.062	2-oxoglutarate^{2-}	48	0.078
Phosphate^{3-}	178	0.041	Hydrogen peroxide	45	0.070
ADP	173	0.026	Pyruvate	38	0.070
NADPH	166	0.021	Glycine	31	0.041
NADP$^+$	166	0.021	S-adenosyl-L-methionine	28	0.063
NAD$^+$	143	0.018	Acetate	26	0.021
Carbon dioxide	139	0.020	S-adenosyl-L-homocysteine	25	0.039
NADH	139	0.017	(6S)-5,6,7,8-tetrahydrofolic acid	25	0.031
AMP	128	0.029	L-glutamine	25	0.025
Coenzyme A	119	0.021	Succinate^{2-}	24	0.029
Dioxygen	116	0.014	Acyl-carrier protein	23	0.013
Ammonium	92	0.011	L-cysteine	22	0.023

[a]The number of metabolites that co-occur in metabolic reactions. [b]The betweenness quantifies the number of paths between any two pairs of metabolites in the network that this one mediates (a global property).

We believe the reconstruction presented here is currently the most comprehensive and consistent stoichiometric representation of yeast metabolism, from which predictive (sub)models, for example for genome-scale flux balance analysis, can be extracted and deployed. Presently, the reconstruction lacks information on effectors, reaction kinetics and parametrization. However, the basic framework of B-Net coupled to SBML models that can easily be populated with such data enables these to be added as they become available, and thus kinetic models that can be directly linked to the genome-scale metabolic network can be built. Some parameters are already available at the System for the Analysis of Biochemical Pathways–Reaction Kinetics (SABIO-RK) website (http://sabio.villa-bosch.de/SABIORK/).

Network reconstruction approaches have developed rapidly in recent years. When they reach the genome scale, they can be viewed as systems-level genome annotations[57]. Genome annotation is produced by a community-driven process to reach a consensus annotation that represents the state of knowledge about the genome of the target organism. Annotations are then updated based on new information and they serve as a common denominator for genome science studies of the target organism. The yeast metabolic reconstruction presented here represents an analogous process for systems biology studies of a target organism. With the successful achievement of the first consensus reconstruction, the systems biology community can look forward to similar two-dimensional annotation jamborees for other organisms.

The metabolite nomenclature proposed here will, we hope, become the standard terminology for metabolic models because the compounds themselves are essentially identical in all species. We believe that the semantically annotated reconstruction provided here will have special utility in a number of areas. First is the basic exploration of metabolic pathways and well-curated connections between gene products. Further, the reconstruction will allow the automated interpretation and visualization of metabolomics data as well as data on metabolic proteins, genes and transcripts. The network can form the basis of phenotype predictions, including product yield, in response to genetic and/or environmental perturbations using a variety of methods, including flux balance analysis and logical approaches[58]. It can also be used in metabolic flux estimation based on isotopomer data[59], for filling gaps in metabolic pathways and for exploring questions related to comparative metabolomics[60] and of metabolic pathway evolution. The widespread use of a consensus starting point will make both the comparison and the integration of such studies considerably easier.

Note added in proof: Nookaew et al.[61] have added useful knowledge of some of the lipid metabolism of baker's yeast.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
The Manchester groups thank the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council (EPSRC) for financial support including for the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org/). The UCSD participants thank the National Institutes of Health for financial support (NIH R01 GM071808). We thank Diane Kelly, Sarah Keating and Norman Paton for many useful discussions. The Jamboree was held under the auspices and with the sponsorship of the Yeast Systems Biology Network (EC Contract: LSHG-CT-2005-018942).

AUTHOR CONTRIBUTIONS
All authors conceived the idea of the consensus reconstruction, the majority were present during the jamboree itself and all contributed to the writing of, and approved, the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Kell, D.B. Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J.* **273**, 873–894 (2006).

2. Arakawa, K., Yamada, Y., Shinoda, K., Nakayama, Y. & Tomita, M. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* **7**, 168 (2006).
3. Palsson, B.Ø. *Systems Biology: Properties of Reconstructed Networks.* (Cambridge University Press, Cambridge; 2006).
4. Duarte, N.C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* **104**, 1777–1782 (2007).
5. Mager, W.H. & Winderickx, J. Yeast as a model for medical and medicinal research. *Trends Pharmacol. Sci.* **26**, 265–273 (2005).
6. Goffeau, A. *et al.* Life With 6000 genes. *Science* **274**, 546–567 (1996).
7. Winzeler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
8. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
9. Yen, K., Gitsham, P., Wishart, J., Oliver, S.G. & Zhang, N. An improved *tet*O promoter replacement system for regulating the expression of yeast genes. *Yeast* **20**, 1255–1262 (2003).
10. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
11. Allen, J.K. *et al.* High-throughput characterisation of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696 (2003).
12. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
13. Castrillo, J.I. *et al.* Growth control of the eukaryote cell: a systems biology study in yeast. *J. Biol.* **6**, 4 (2007).
14. Delneri, D. *et al.* Identification and characterization of high-flux-control genes of yeast through competition analyses in continuous cultures. *Nat. Genet.* **40**, 113–117 (2008).
15. Wu, J., Zhang, N., Hayes, A., Panoutsopoulou, K. & Oliver, S.G. Global analysis of nutrient control of gene expression in *Saccharomyces cerevisiae* during growth and starvation. *Proc. Natl. Acad. Sci. USA* **101**, 3148–3153 (2004).
16. Oliver, S. A network approach to the systematic analysis of gene function. *Trends Genet.* **12**, 241–242 (1996).
17. Suter, B., Auerbach, D. & Stagljar, I. Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques* **40**, 625–644 (2006).
18. Förster, J., Famili, I., Fu, P., Palsson, B.Ø. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253 (2003).
19. Duarte, N.C., Herrgard, M.J. & Palsson, B.Ø. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309 (2004).
20. Kuepfer, L., Sauer, U. & Blank, L.M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae. Genome Res.* **15**, 1421–1430 (2005).
21. Caspi, R. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **34**, D511–D516 (2006).
22. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
23. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**, 1509–1515 (2005).
24. Çakir, T. *et al.* Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol. Syst. Biol.* **2**, 50 (2006).
25. Kümmel, A., Panke, S. & Heinemann, M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* **2**, 2006.0034 (2006).
26. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.Ø. An expanded genome-scale model of *Escherichia coli* K12 (iJR904 GSM/GPR). *Genome Biol* **4**, R54 (2003).
27. Förster, J., Famili, I., Palsson, B.Ø. & Nielsen, J. Large-scale evaluation of *in silico* deletions in *Saccharomyces cerevisiae. OMICS* **7**, 193–202 (2003).
28. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357 (2006).
29. Nash, R. *et al.* Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.* **35**, D468–D471 (2007).
30. Caspi, R. *et al.* The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **36**, D623–D631 (2008).
31. Li, X.J. *et al.* in Metabolic profiling: its role in biomarker discovery and gene function analysis. (eds. Harrigan, G.G. & Goodacre, R.) 293–309 (Kluwer Academic Publishers, Boston, 2003).
32. Goble, C. & Wroe, C. The Montagues and the Capulets. *Comp. Funct. Genomics* **5**, 623–632 (2004).
33. Ananiadou, S., Kell, D.B. & Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol.* **24**, 571–579 (2006).
34. Poolman, M.G., Bonde, B.K., Gevorgyan, A., Patel, H.H. & Fell, D.A. Challenges to be faced in the reconstruction of metabolic networks from public databases. *Syst. Biol. (Stevenage)* **153**, 379–384 (2006).
35. Spasić, I. *et al.* Facilitating the development of controlled vocabularies for metabolomics with text mining. *BMC Bioinformatics* **9**, S5 (2008).
36. Williams, A.J. Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* **13**, 502–506 (2008).
37. Williams, A.J. A perspective of publicly accessible/open-access chemistry databases. *Drug Discov. Today* **13**, 495–501 (2008).
38. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3**, 135 (2007).
39. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* **33**, D46–D53 (2005).
40. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).
41. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
42. Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S. & Zhang, Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **3**, 1832–1834 (2005).
43. Wishart, D.S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
44. The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* **36**, D190–D195 (2008).
45. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**, D527–D532 (2007).
47. Barabási, A.-L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
48. Wagner, A. & Fell, D.A. The small world inside large metabolic networks. *Proc. R. Soc. Lond., B, Biol. Sci.* **268**, 1803–1810 (2001).
49. Hoops, S. *et al.* COPASI: a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067–3074 (2006).
50. Vallabhajosyula, R.R., Chickarmane, V. & Sauro, H.M. Conservation analysis of large biochemical networks. *Bioinformatics* **22**, 346–353 (2006).
51. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
52. Funahashi, A., Tanimura, N., Morohashi, M. & Kitano, H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* **1**, 159–162 (2003).
53. Li, P., Oinn, T., Soiland, S. & Kell, D.B. Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics* **24**, 287–289 (2008).
54. Bornstein, B.J., Keating, S.M., Jouraku, A. & Hucka, M. LibSBML: an API library for SBML. *Bioinformatics* **24**, 880–881 (2008).
55. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few* (Abacus, London, 2004).
56. Tapscott, D. & Williams, A. *Wikinomics: How Mass Collaboration Changes Everything* (New Paradigm, Toronto, 2007).
57. Palsson, B. Two-dimensional annotation of genomes. *Nat. Biotechnol.* **22**, 1218–1219 (2004).
58. Whelan, K.E. & King, R.D. Using a logical model to predict the growth of yeast. *BMC Bioinformatics* **9**, 97 (2008).
59. Blank, L.M., Kuepfer, L. & Sauer, U. Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* **6**, R49 (2005).
60. Kell, D.B. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov. Today* **11**, 1085–1092 (2006).
61. Nookaew, I. *et al.* The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC Syst. Biol.* **2**, 71 (2008).

[1]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0412, USA. [2]School of Computer Science, The University of Manchester, Oxford Rd., Manchester M13 9PL, UK. [3]The Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St., Manchester M1 7DN, UK. [4]School of Chemistry, The University of Manchester, Manchester M13 9PL, UK. [5]Department of Chemical Engineering, Boğaziçi University, Bebek 34342, Istanbul, Turkey. [6]VTT Biotechnology Espoo, PO Box 1500, FIN-02044, Finland. [7]School of Chemical Engineering and Analytical Science, The University of Manchester, UK. [8]Max-Planck-Institut für Molekulare Genetik, Ihnestrasse 73, 14195 Berlin, Germany. [9]Institut für Molekulare Systembiologie, ETH Zurich Wolfgang-Pauli-Str. 16, 8093 Zürich, Switzerland. [10]Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125, USA. [11]Computational Neurobiology, EMBL-EBI, Wellcome-Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [12]Center for Microbial Biotechnology, Department of Systems Biology, Technical University of Denmark, Building 223, DK-2800 Kgs. Lyngby, Denmark. [13]School of Mathematics, The University of Manchester, Manchester M13 9PL, UK. [14]The Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA. [15]Department of Molecular Cell Physiology, Vrije Universiteit, de Boelelaan 1085, 1081 HV Amsterdam, The Netherlands. [16]Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, UK. [17]Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, Virginia 24061, USA. [18]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [19]Present addresses: Synthetic Genomics, Inc., 11149 N. Torrey Pines Rd., La Jolla, California 92037, USA (M.J.H.) and Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden (D.P.). [20]These authors contributed equally to this work. Correspondence should be addressed to D.B.K. (dbk@manchester.ac.uk).

Publication 2

Further developments towards a genome-scale metabolic model of yeast.

Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, **Swainston N**, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD, Mendes P.

BMC Syst Biol. 2010, **4**, 145.

Dobson et al. BMC Systems Biology 2010, 4:145
http://www.biomedcentral.com/1752-0509/4/145

BMC Systems Biology

RESEARCH ARTICLE Open Access

Further developments towards a genome-scale metabolic model of yeast

Paul D Dobson[1†], Kieran Smallbone[2,3*†], Daniel Jameson[2,4], Evangelos Simeonidis[2,5], Karin Lanthaler[1,2], Pınar Pir[6], Chuan Lu[7], Neil Swainston[2,4], Warwick B Dunn[1,2], Paul Fisher[4], Duncan Hull[1], Marie Brown[1], Olusegun Oshota[2,5,8], Natalie J Stanford[2,5,8], Douglas B Kell[1], Ross D King[7], Stephen G Oliver[6], Robert D Stevens[4], Pedro Mendes[2,4,9]

Abstract

Background: To date, several genome-scale network reconstructions have been used to describe the metabolism of the yeast *Saccharomyces cerevisiae*, each differing in scope and content. The recent community-driven reconstruction, while rigorously evidenced and well annotated, under-represented metabolite transport, lipid metabolism and other pathways, and was not amenable to constraint-based analyses because of lack of pathway connectivity.

Results: We have expanded the yeast network reconstruction to incorporate many new reactions from the literature and represented these in a well-annotated and standards-compliant manner. The new reconstruction comprises 1102 unique metabolic reactions involving 924 unique metabolites - significantly larger in scope than any previous reconstruction. The representation of lipid metabolism in particular has improved, with 234 out of 268 enzymes linked to lipid metabolism now present in at least one reaction. Connectivity is emphatically improved, with more than 90% of metabolites now reachable from the growth medium constituents. The present updates allow constraint-based analyses to be performed; viability predictions of single knockouts are comparable to results from *in vivo* experiments and to those of previous reconstructions.

Conclusions: We report the development of the most complete reconstruction of yeast metabolism to date that is based upon reliable literature evidence and richly annotated according to MIRIAM standards. The reconstruction is available in the Systems Biology Markup Language (SBML) and via a publicly accessible database http://www.comp-sys-bio.org/yeastnet/.

Background

A central goal of integrative systems biology is the accurate representation of molecular interaction networks. Ultimately, such networks can be used to underpin mathematical models, consisting of stochastic or ordinary differential equations that permit the simulation of biological behaviour. The first step in generating such models is constructing a network of biochemical reactions and interactions between molecular components of the system to form a qualitative (unparameterised) model. Several groups have reconstructed the metabolic network of baker's yeast from genomic and literature data [1-3]. Variation in the approaches used, and contradictory interpretations of the available literature, mean that most reconstructions differ considerably. To resolve these problems, a cohort of the yeast systems biology community collaborated to create a consensus reconstruction. In April 2007, a large focused meeting brought together experts from various groups and disciplines in order to resolve discrepancies between the various reactions and metabolites described by other available reconstructions and form a consensus. The resultant reconstruction [4], subsequently referred to as "Yeast 1.0", removed the ambiguities inherent in its predecessors through the use of principled and computer-readable annotations. Whilst previous reconstructions had defined entities using subjective names, which lacked precision and resulted in ambiguities, Yeast 1.0 directly referenced chemical and protein descriptions to persistent databases or used standardised, database-independent, computer-readable

* Correspondence: kieran.smallbone@manchester.ac.uk
† Contributed equally
[2]Manchester Centre for Integrative Systems Biology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK
Full list of author information is available at the end of the article

representations. This removed the ambiguities and allowed the new reconstruction to be used effectively as the basis for automated analyses.

A limitation of Yeast 1.0 came about through the very generation of the consensus; the network became considerably fragmented as reactions that could not be readily annotated (due to the presence of structural ambiguities) were removed. This led to underrepresentation of a number of pathways, particularly those involved in lipid biosynthesis. Since Yeast 1.0, many improvements have been made to the reconstruction. The latest release, described here, is considerably larger (in terms of numbers of metabolites and reactions), of higher quality (by reference to literature evidence), exhibits greater coverage of known metabolic enzymes, and is better connected than all previous efforts.

The reconstruction is described and made available in Systems Biology Markup Language (SBML) [5], an established community XML format for the mark-up of biochemical models. With the introduction of SBML Level 2, specific model entities, such as species or reactions, can be annotated using ontological terms. These annotations, encoded using the resource description framework (RDF) [6], provide the facility to assign definitive terms to individual components, allowing the software to identify such components unambiguously and thus link model components to existing data resources [7]. Minimum Information Requested in the Annotation of Models (MIRIAM) [8] -compliant annotations have been used to identify components unambiguously by associating them with one or more terms from publicly available databases registered in MIRIAM Resources [9]. An example of such an annotation is presented in Figure 1, where an enzyme is identified by MIRIAM-compliant references to the UniProt [10], SGD [11], and PubMed [12] databases. Metabolites are annotated with reference to the ChEBI (Chemical Entities of Biological Interest) database [13]. Whilst SBML is the primary format for dissemination of the reconstruction, we also make the reconstruction available in an online database [14], B-Net, that enables easy searching of the content. B-Net [15] is able to represent all of the SBML features utilised in the current reconstruction. Searches can be performed using synonyms and the user is also able to navigate through the network from any point (e.g. a metabolite, reaction or enzyme) to its connected neighbours. Query results can also be exported in SBML and this is an effective mechanism to extract subsets of the entire model in this exchange format.

Results and Discussion

Improvements in the representation of yeast metabolism in this release as compared to Yeast 1.0 primarily consist of its enhanced representation of lipid metabolism and greater connectivity, thereby permitting constraint-based flux analyses. Many of the extensions to Yeast 1.0 are reactions garnered from the literature, which are entirely novel to any genome-wide yeast metabolic reconstruction. Data were also incorporated, when backed up by traceable evidence, from two other reconstructions: iMM904 [16] and iIN800 [17]. The resulting consensus network (reported in Additional File 1) consists, in decompartmentalised form, of 1102 metabolic reactions involving 924 metabolites and 924 proteins (Table 1) and is therewith larger in scope than any previous reconstruction.

Careful curation does not simply involve increasing the scope of the reconstruction. Indeed, 32 enzymes from Yeast 1.0 were considered insufficiently evidenced and have been removed, whilst a number of metabolites were relocalised to a different compartment. A typical example of an enzyme removed from the reconstruction is Gpm2p; whilst a homologue of Gpm1p, its phosphoglycerate mutase activity could not be evidenced and may be non-functional [18]. Four reconstructions are compared in Figure 2 in terms of enzymes present. In addition to the 32 enzymes removed, the reactions of a further 37 enzymes from iMM904 and iIN800 have not been added for lack of supporting evidence. In total, the new reconstruction considers 124 more enzymes than its predecessor, with half of these (61) being retrieved manually from the literature and therefore new to all reconstructions.

Lipid metabolism

The correct and complete representation of lipid metabolism is important, not only to meet the ultimate goal of genome-scale coverage, but also because understanding and engineering lipid metabolism through systems and synthetic biology is likely to play a major role in the replacement of fossil energy sources and chemical feedstocks with biofuels and bioplastics [19]. In Yeast 1.0, lipid metabolism was poorly captured. To move towards a better representation, the literature, database annotations and homology relationships were used to identify the set of lipid-related yeast enzymes. Homology with mouse and human enzymes reported in LipidMaps [20], and with enzymes from all organisms reported in KEGG lipid pathways [21], indicated lipid enzymes in yeast (homology relationships predefined by Ensembl [22]). Further enzymes were added to the set manually by examination of SGD and Ensembl annotations. A total of 268 yeast enzymes were identified as likely to be part of lipid metabolism. Although the boundaries of this set are unavoidably subjective, it appears to capture the majority of lipid-related genes in yeast.

With reference to this set of lipid enzymes, the iIN800 reconstruction of Nookaew *et al.* improved upon the original community reconstruction (Yeast 1.0) by increasing set coverage from 48% to 62% (with at least one reaction being associated with each enzyme). In the

```
<speciesType metaid="_metaE_1" id="E_1" name="YBL022C">
  <annotation>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
      <rdf:Description rdf:about="#_metaE_1">
        <bqbiol:isDescribedBy>
          <rdf:Bag>
            <rdf:li rdf:resource="urn:miriam:pubmed:8276800"/>
          </rdf:Bag>
        </bqbiol:isDescribedBy>
        <bqbiol:is>
          <rdf:Bag>
            <rdf:li rdf:resource="urn:miriam:uniprot:P36775"/>
          </rdf:Bag>
        </bqbiol:is>
        <bqbiol:isEncodedBy>
          <rdf:Bag>
            <rdf:li rdf:resource="urn:miriam:sgd:S000000118"/>
          </rdf:Bag>
        </bqbiol:isEncodedBy>
      </rdf:Description>
    </rdf:RDF>
  </annotation>
</speciesType>
```

Figure 1 SBML example. Simplified example of MIRIAM-compliant SBML, whereby an enzyme is annotated with reference to the databases UniProt, SGD and PubMed, respectively.

present release set coverage has further improved to 87%. Coverage of the lipid enzyme set by the various reconstructions is summarised in Figure 3. From iIN800 and iMM904, 56 lipid enzymes were added to Yeast 1.0, while three enzymes from these sources were not added. The current reconstruction describes activities for 49 enzymes that no other reconstruction has ever considered. Combining these, the reconstruction extends the Yeast 1.0 description of lipid metabolism by a total of 105 new enzymes, extends iMM904 by 59 enzymes, and iIN800 by 70 enzymes. This is by far the most comprehensive reconstruction of yeast lipid metabolism to date.

The 34 remaining lipid enzymes (in figure 3 these are 31 not found in any reconstruction, plus three found in both iMM904 and iIN800) from the set are either too poorly characterised functionally to be included or cannot be represented within the current description of the cell's compartmentalisation. Flippases, for example, require a more detailed description of membrane faces to capture their role in membrane asymmetry. Improving compartmental representation will be a goal for future releases.

Table 1 Reconstruction scope

	iMM904	iIN800	Yeast 1.0	Yeast 4.0	change (%)
reactions	1050	907	962	1102	14.6
metabolites	872	812	813	924	13.7
proteins	904	707	832	924	11.0
compartments	8	4	15	16	6.7

Comparison of the scope of reconstructions (Yeast 4.0 being the version number of the current reconstruction). Metabolites and reactions in different intracellular compartments are considered one, as are reactions with the same stoichiometry (isoenzymatic). This renders reconstructions with differing granularity comparable.

Connectivity

Structural improvement was a major focus of the advancements made to the reconstruction by identifying

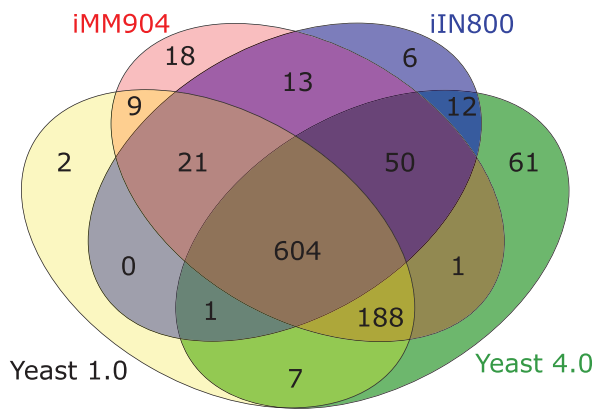

Figure 2 Comparison of reconstructions in terms of enzymes present. The reconstruction presented here contains 124 more enzymes than Yeast 1.0, 61 of which have not been considered by any of the other reconstructions. Yeast 1.0 was also improved upon through better curation leading to the removal of (2 + 9 + 21 =) 32 enzymes. A further (6 + 13 + 18 =) 37 enzymes from iMM908 and iIM800 were not added to the reconstruction.

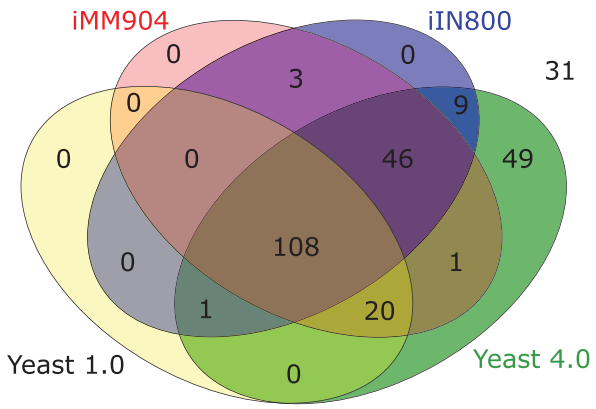

Figure 3 Comparison of the coverage of lipid metabolism enzymes by the different reconstructions. At least one reaction in a reconstruction is catalyzed by each enzyme. On top of extending Yeast 1.0 by (1 + 9 + 46 =) 56 enzymes from iMM904 and iIN800, a further 49 enzymes uniquely appear in this latest reconstruction. Three reactions common to iMM904 and iIN800, plus 31 others, have not been incorporated for lack of evidence.

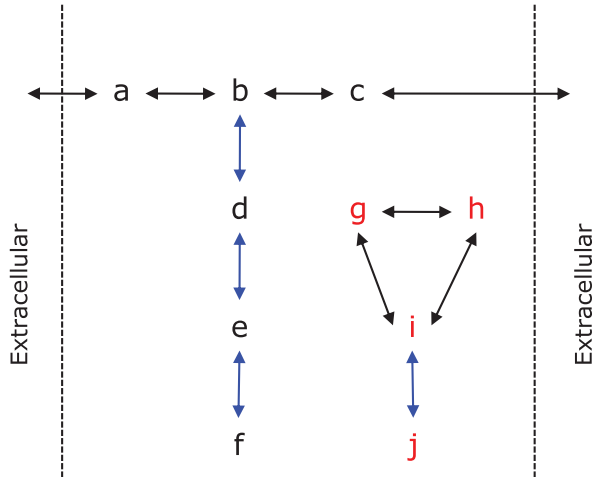

Figure 4 Visualisation of connectivity analysis. Metabolites that are unreachable (in red) were identified with a graphical analysis, by locating metabolites that are disconnected from the extracellular medium. Flux variability analysis identified reactions that must have zero flux (in blue) because they lead to dead-end metabolites.

and rectifying unconnected regions of the network. Two measures were used to describe connectivity. First, we identified clusters of unreachable metabolites; that is, clusters of metabolites that are disconnected from the extracellular medium, in a graph-theoretic sense, and thus cannot ever be produced by the reaction network. Secondly, we used flux variability analysis [23] to identify reactions that, by mass balancing, must have zero flux, for example because of dead-end metabolites (products that are not the substrates of another reaction). Led by these analyses, which are explained graphically in Figure 4, we looked for literature evidence describing these missing elements of our network. By targeting unreachable clusters and those reactions whose reconnection has the most influence on the network's connectivity, we maximised the impact of literature curation on modelling. By both measures, the present release improves both upon the previous release and particularly upon iMM904 and iIN800 (Table 2). More than 90% of metabolites can be reached from the extracellular medium and only 12.7% of reactions must have zero flux.

Our approach towards structural improvement is also an example of the iterative "cycle of knowledge" approach [24], where the model is first used to guide biological research and can subsequently be updated and improved as specific new knowledge becomes available. In this case the iteration consisted of discovery and collation of experimental evidence previously obtained but which had never been identified in this context. Such discovery of knowledge was informed by the previous models and was unlikely to have happened in their absence.

Constraint-based analysis

New reconstructions are often validated through constraint-based approaches like Flux Balance Analysis (FBA) [25] to assess their ability to predict experimental results. While there is clear utility in deploying such methods to explore biochemical capacity, using improved agreement with experimental observations to determine whether the reconstruction is, in some sense, 'better' than previous efforts is potentially misleading. In the current release, non-inferred reactions are supported by evidence from the literature and it is in this sense that the reconstruction is validated and improved. That said, the updates improved the connectivity considerably and together with the inclusion of a reaction describing biomass composition now allows FBA to be performed. The availability of the model in SBML means that it is accessible through many generic and systems-biology-specific software packages, including the COBRA (COnstraint-Based Reconstruction and Analysis) toolbox [26].

The model was used to predict single knockout viability through flux balance analysis (FBA). Growth conditions

Table 2 Network connectivity

	iMM904	iIN800	Yeast 1.0	Yeast 4.0
intracellular metabolites	708	681	658	758
unreachable	440	468	108	75
%	62.2	68.7	16.4	9.9
metabolic reactions	1050	907	962	1102
zero flux	225	282	153	140
%	21.4	31.1	15.9	12.7

As in Table 1, decompartmentalised models were used to generate these data.

exactly followed those set out in iMM904, namely a glucose-limited minimal medium. Cellular biomass was defined as in iIN800 (carbon-limited version), due to its high level of detail regarding lipid composition. As the reaction producing biomass does not represent a real metabolic process it is semantically annotated as such using SBO (Systems Biology Ontology) [27] identifiers and GO (Gene Ontology) [28] evidence codes to ensure this distinction is maintained (therefore allowing one to easily remove this reaction based on its annotation). Simulations were performed using COBRA (which is reliant on libSBML [29] and the GNU linear programming kit [30]). The simulation predictions were compared to a list of lethal gene knockouts. This list was generated by considering results from viability experiments under both rich [31] and glucose minimal growth medium conditions [32]. Results demonstrate similar performance to that of previous reconstructions in terms of the accuracy of prediction of single gene knockout viability (Table 3).

Closer inspection of predictions reveals that relatively subtle network variations often underlie prediction differences. Four experimentally lethal knockouts were not initially predicted as such by the new reconstruction, but are correctly predicted using iMM904. Three of these genes encode enzymes that are essential to riboflavin biosynthesis. The capacity of iMM904 to predict lethality correctly is due to its biomass definition including a small contribution from riboflavin, whereas this was not part of the initial iIN800 or current network's biomass definition. Subsequent addition of riboflavin to the (empirical) biomass description has resolved these differences. Note that this is not therefore a reflection of the quality of the underlying network but only of the empirical biomass estimation, which is itself dependent on the growth conditions.

In places, the added richness of the new reconstruction combines with certain known limitations to defeat total agreement with experiment. An example is seen by knocking out the *acs2* gene, encoding acetyl-coA synthetase (Acs2p). By experiment this should be lethal, yet in the current network the cytoplasmic reaction is also catalysed by Acs1p, consistent with experimental data [33]. When the Acs2p-catalysed reaction is eliminated, flux simply re-routes through the Acs1p reaction. Importantly, it is only the fortuitous incompleteness of iMM904, lacking the cytosolic Acs1 isozyme that reveals the inviability of the *acs2* knockout. The proper basis of the inviability of the *acs2* mutant is that *ACS1* is transcriptionally repressed in the high glucose conditions of viability experiments and so is unable to compensate for the loss of *ACS2* [34]. Transcriptional control is not captured in the metabolic network and thus cannot be captured in metabolic reconstructions of this type.

Both these examples highlight the caution required when using approaches such as FBA to validate reconstructions. The added detail in the present network can naturally lead to an increase in false positive outcomes: *in silico* knockouts that are overcome by alternative routings in the network but are actually lethal *in vivo*. This is, however, tempered by a decrease in false negative outcomes (i.e. knockouts that appear lethal computationally but are viable *in vivo*, as presented in Table 3).

Table 3 Gene knockout analysis

	iMM904	iIN800	Yeast 4.0
number of genes	904	707	924
true positive (%)	75.0	69.7	74.8
true negative (%)	5.1	6.9	5.3
false positive (%)	9.3	10.6	11.1
false negative (%)	10.6	12.7	8.8

Results of *in silico* viability prediction of deletion strains of *S. cerevisiae*. "Positive" and "negative" refer to the ability and inability to grow, respectively. Following earlier studies, the knockout simulation was conducted in a glucose-limited minimal medium, and compared to experimental knockout data [30,31].

Uncharacterised enzymes

Despite the much-increased coverage of the current reconstruction, 451 genes probably encode metabolic enzymes that still have no associated reaction (Additional file 2). For the majority of these, very little is known about their function and further characterisation is required. From the viewpoint of furthering systems biology reconstruction efforts, these enzymes are important targets for reductionist molecular biology studies, including, for instance, systematic analyses using the Robot Scientist approach [35]. Their listing here is a motivation for further iterations on the cycle of knowledge.

Conclusions

The development of high quality, well annotated, genome-scale, metabolic networks is an ambitious, challenging, but necessary step towards the realisation of integrative systems biology. While networks predicted through bioinformatics approaches are useful, particularly for the extension of systems biology approaches to less well-studied organisms, reconstructions built upon solid biochemical evidence provide a gold standard upon which predictions can be reliably based. For metabolic reconstructions, where the goal is to capture maximally our current understanding of metabolism, these problems are primarily of data integration and quality. It has proven essential to involve the extended systems biology and yeast communities in this process, both to establish the mechanisms and structures for acquiring and representing information, and also to tap into expert knowledge from the various sub-disciplines of biology and biochemistry. In the recent very large-scale

reconstruction of the yeast molecular interaction network by Aho *et al.* [36], genomic, transcriptomic, proteomic and metabolomic data were integrated. These authors note that incorporating the higher quality data of Yeast 1.0 (and therefore even more of this contribution) would considerably improve their reconstruction over the metabolic information extracted from KEGG, and also that standards compliance is essential to this integration task.

Yeast 1.0 set standards and amalgamated existing networks, enhancing annotation and removing less reliable data. In this latest reconstruction, we have made significant headway on the process of filling gaps in the network. There is still some way to go before realising the goal of at least one reaction for each putative metabolic enzyme and, if one also considers enzyme promiscuity [37,38], even this will represent an incomplete picture of metabolism. This latest reconstruction is a considerable improvement on previous releases, particularly in describing lipid metabolism and addressing gaps in the original reconstruction that hindered modelling efforts. Information from other reconstructions since Yeast 1.0 has been incorporated, although not indiscriminately, and very many reactions not found in other reconstructions have been garnered from the literature. It is considerably larger than all previous efforts, while maintaining compliance with community-defined standards.

While Yeast 1.0 represented a major advance, particularly through the definition of standards and by the involvement of the wider yeast community, a major flaw was that it was not amenable to constraint-based analysis. The current reconstruction rectifies this, mostly by filling in gaps but also by inclusion of an appropriately annotated "biomass" reaction, without compromising the strict evidence requirements of its predecessor. When compared to experimental knockout data, this reconstruction did not identify certain lethal knockouts that other yeast reconstructions correctly predicted, but proves better than them in recognising viable deletions. This is a direct result of the richness of the model; as with the example of the acetyl-coA synthetases (above), addition of isoenzymes of specific reactions that do not exist in earlier reconstructions can reduce the predictive power of the model. Nonetheless, such enzymes are included due to literature support. This reconstruction continues the shifting focus, started with the consensus model Yeast 1.0, toward realistic representation and proof-based selection of reactions, rather than creating a reconstruction with simulation in mind. Reactions with a lower level of confidence (e.g. biomass definition) are characterised with specialised evidence codes and SBO terms, allowing the easy extraction of subsets of the network from the SBML code for specific purposes.

To facilitate further improvements, we encourage the community to provide information and/or corrections to the current release. We have set up a dedicated point-of-contact to this end network.reconstruction@manchester.ac.uk. We also highlight gaps in the network that cannot be resolved from current literature, as well as the little-studied enzymes for which we have not yet identified any function (see Additional File 2). These represent potentially important research opportunities for the community and we welcome efforts towards an improved understanding of their functions.

Additional material

Additional file 1: Yeast SBML files. ZIP file containing the latest reconstruction in SBML format. The metabolic network reconstruction is described using MIRIAM-compliant SBML, compatible with many Systems Biology software packages, including the COBRA toolbox. The model is also available in decompartmentalised form, and in an old SBML format (level 2, version 1) for backward compatibility.

Additional file 2: Poorly characterised genes. Excel spreadsheet. The network is built upon intensive literature mining to identify reactions. Many genes still do not have detailed literature describing the functions of their products, yet (by what little is known or through sequence analysis) they appear likely to be involved in metabolism. The attached list describes these genes.

Acknowledgements

The Manchester groups thank the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council (EPSRC) for financial support (grants BB/C008219/1 and BB/F006012/1). The Cambridge group acknowledges BBSRC grant BB/C505140/2. The Manchester, Aberystwyth and Cambridge groups all acknowledge support from the European Union FP7 project UNICELLSYS (Grant agreement no.: 201142) and from SysMO (MOSES). We thank Mike Hucka for advice on formatting SBML annotations, Rasmus Ågren for providing the iIN800 reconstruction and Steve Turner for help with ChEBI submissions. This is a contribution from the Manchester Centre for Integrative Systems Biology and the Cambridge Systems Biology Centre.

Author details

[1]School of Chemistry, The University of Manchester, Manchester M13 9PL, UK. [2]Manchester Centre for Integrative Systems Biology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK. [3]School of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK. [4]School of Computer Science, Kilburn Building, The University of Manchester, Oxford Road, Manchester M13 9PL, UK. [5]School of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK. [6]Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK. [7]Department of Computer Science, Aberystwyth University, SY23 3DB, UK. [8]Doctoral Training Centre for Integrative Systems Biology, The University of Manchester. [9]Virginia Bioinformatics Institute, Virginia Tech, Washington Street 0477, Virginia 24061, USA.

Authors' contributions

PDD, KS, DJ, ES, KL, PP, NS, WBD, DH, MB, OO, NJS and PM contributed to literature curation to identify new reactions. KS and NS prepared and curated the SBML. PF collated relevant literature for curation. PDD, KS, DJ, ES, DBK and PM wrote the manuscript. CL, DBK, RDK, SGO, RDS and PM supervised work and/or contributed to discussions. All authors read, improved, and approved the final manuscript.

Received: 2 June 2010 Accepted: 28 October 2010
Published: 28 October 2010

References

1. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network.** *Genome Research* 2003, **13(2)**:244-253.
2. Duarte NC, Herrgård MJ, Palsson BØ: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Research* 2004, **14(7)**:1298-1309.
3. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in Saccharomyces cerevisiae.** *Genome Research* 2005, **15(10)**:1421-1430.
4. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, *et al*: **A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.** *Nature Biotechnology* 2008, **26(10)**:1155-1160.
5. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19(4)**:524-531.
6. Wang XS, Gorlitsky R, Almeida JS: **From XML to RDF: how semantic web technologies will change the design of 'omic' standards.** *Nature Biotechnology* 2005, **23(9)**:1099-1103.
7. Kell DB, Mendes P: **The markup is the model: reasoning about systems biology models in the Semantic Web era.** *Journal of Theoretical Biology* 2008, **252(3)**:538-543.
8. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, *et al*: **Minimum information requested in the annotation of biochemical models (MIRIAM).** *Nature Biotechnology* 2005, **23(12)**:1509-1515.
9. Laibe C, Le Novere N: **MIRIAM resources: tools to generate and resolve robust cross-references in Systems Biology.** *BMC Systems Biology* 2007, **1**:58.
10. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, *et al*: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Research* 2010, **38**:D142-D148.
11. Weng S, Dong Q, Balakrishnan R, Christie K, Costanzo M, Dolinski K, Dwight SS, Engel S, Fisk DG, Hong E, *et al*: **Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins.** *Nucleic Acids Research* 2003, **31(1)**:216-218.
12. **PubMed.** [http://www.ncbi.nlm.nih.gov/pubmed/].
13. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical Entities of Biological Interest: An update.** *Nucleic Acids Research* 2009, **38**:D249-254.
14. **YeastNet: A consensus reconstruction of yeast metabolism.** [http://www.comp-sys-bio.org/yeastnet/].
15. **B-Net: A schema for representing detailed biochemical knowledge.** [http://mendes.vbi.vt.edu/tiki-index.php?page=B-Net].
16. Mo ML, Palsson BØ, Herrgård MJ: **Connecting extracellular metabolomic measurements to intracellular flux states in yeast.** *BMC Systems Biology* 2009, **3**:37.
17. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, Nielsen J, Bhumiratana S: **The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism.** *BMC Systems Biology* 2008, **2**:71.
18. Heinisch JJ, Müller S, Schlüter E, Jacoby J, Rodicio R: **Investigation of two yeast genes encoding putative isoenzymes of phosphoglycerate mutase.** *Yeast* 1998, **14(3)**:203-213.
19. Ratledge C, Cohen Z: **Microbial and algal oils: Do they have a future for biodiesel or as commodity oils?** *Lipid Technology* 2008, **20(7)**:155-160.
20. Fahy E, Sud M, Cotter D, Subramaniam S: **LIPID MAPS online tools for lipid research.** *Nucleic Acids Research* 2007, **35**:W606-612.
21. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Research* 2006, **34**:D354-D357.
22. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Research* 2009, **37**:D690-D697.
23. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.** *Metabolic Engineering* 2003, **5(4)**:264-276.
24. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.** *Bioessays* 2004, **26(1)**:99-105.
25. Kauffman KJ, Prakash P, Edwards JS: **Advances in flux balance analysis.** *Current Opinion in Biotechnology* 2003, **14(5)**:491-496.
26. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgård MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nature Protocols* 2007, **2(3)**:727-738.
27. Le Novère N, Courtot M, Laibe C: **Adding semantics in kinetics models of biochemical pathways.** *Proceedings of the 2nd International Symposium on experimental standard conditions of enzyme characterizations: 2006* Rüdesheim, Germany Beilstein Institut; 2006, 137-153.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25(1)**:25-29.
29. Bornstein BJ, Keating SM, Jouraku A, Hucka M: **LibSBML: An API library for SBML.** *Bioinformatics* 2008, **24(6)**:880-881.
30. Makhorin A: **GNU Linear Programming Kit.** Moscow: Moscow Aviation Institute; 2001.
31. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, *et al*: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418(6896)**:387-391.
32. Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segrè D: **Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions.** *Genome Biology* 2008, **9(9)**:R140.
33. **SGD project: ACS1/YAL054C.** [http://www.yeastgenome.org/cgi-bin/locus.fpl?dbid=S000000050].
34. van den Berg MA, de Jong-Gubbels P, Kortland CJ, van Dijken JP, Pronk JT, Steensma HY: **The two acetyl-coenzyme A synthetases of Saccharomyces cerevisiae differ with respect to kinetic properties and transcriptional regulation.** *Journal of Biological Chemistry* 1996, **271(46)**:28953-28959.
35. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, *et al*: **The Automation of Science.** *Science* 2009, **324(5923)**:85-89.
36. Aho T, Almusa H, Matilainen J, Larjo A, Ruusuvuori P, Aho KL, Wilhelm T, Lähdesmäki H, Beyer A, Harju M, *et al*: **Reconstruction and validation of RefRec: a global model for the yeast molecular interaction network.** *PLoS ONE* **5(5)**:e10662.
37. Hult K, Berglund P: **Enzyme promiscuity: mechanism and applications.** *Trends in Biotechnology* 2007, **25(5)**:231-238.
38. Nobeli I, Favia AD, Thornton JM: **Protein promiscuity and its implications for biotechnology.** *Nature Biotechnology* 2009, **27(2)**:157-167.

doi:10.1186/1752-0509-4-145
Cite this article as: Dobson *et al.*: Further developments towards a genome-scale metabolic model of yeast. *BMC Systems Biology* 2010 **4**:145.

Publication 3

libAnnotationSBML: a library for exploiting SBML annotations.

Swainston N, Mendes P.

Bioinformatics. 2009, **25**, 2292–3.

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 17 2009, pages 2292–2293
doi:10.1093/bioinformatics/btp392

Systems biology

libAnnotationSBML: a library for exploiting SBML annotations

Neil Swainston[1,*] and Pedro Mendes[1,2]

[1]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK and [2]Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, VA 24061, USA

Received on March 20, 2009; revised on May 22, 2009; accepted on June 21, 2009

Advance Access publication June 26, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The Systems Biology Markup Language (SBML) is an established community XML format for the markup of biochemical models. With the introduction of SBML level 2 version 3, specific model entities, such as species or reactions, can now be annotated using ontological terms. These annotations, which are encoded using the resource description framework (RDF), provide the facility to specify definite terms to individual components, allowing software to unambiguously identify such components and thus link the models to existing data resources.

libSBML is an application programming interface library for the manipulation of SBML files. While libSBML provides the facilities for reading and writing such annotations from and to models, it is beyond the scope of libSBML to provide interpretation of these terms. The libAnnotationSBML library introduced here acts as a layer on top of libSBML linking SBML annotations to the web services that describe these ontological terms. Two applications that use this library are described: SbmlSynonymExtractor finds name synonyms of SBML model entities and SbmlReactionBalancer checks SBML files to determine whether specifed reactions are elementally balanced.

Availability: http://mcisb.sourceforge.net/

Contact: neil.swainston@manchester.ac.uk

1 INTRODUCTION

The minimum information requested in the annotation of biochemical models (MIRIAM; Le Novère *et al.*, 2005) defines guidelines for annotation of biochemical models. The annotation of models with the MIRIAM standard provides a number of significant advantages in the development of computational tools and applications that can reason over them (Kell and Mendes, 2008).

An example is the task of comparing or merging two biochemical models. Before the introduction of MIRIAM, individual components of SBML models (Hucka *et al.*, 2003) were identified solely by free-text, human-readable, name attributes, often resulting in equivalent components being named differently in different models. As naming conventions are non-standard, it is impossible to definitively match these components computationally, and the process of model merging then requires human input to resolve ambiguities. Providing MIRIAM-compliant annotations allows a component to be unambiguously identified by associating it with one or more terms from publicly available databases such as ChEBI (Degtyarenko *et al.*, 2008) or KEGG (Kanehisa *et al.*, 2000) (Fig. 1).

```
<species id="species1">
  <annotation>
  <RDF>
  <li resource="urn:miriam:obo.chebi:CHEBI%3A4167"/>
  </RDF>
  </annotation>
</species>

<species id="species2">
  <annotation>
  <RDF>
  <li resource="urn:miriam:kegg.compound:C00031"/>
  </RDF>
  </annotation>
</species>
```

Fig. 1. Simplified example of MIRIAM-compliant SBML species elements, annotated with ChEBI and KEGG terms, respectively.

2 FEATURES

The species elements in Figure 1 are both annotated with MIRIAM-compliant terms. libSBML (Bornstein *et al.*, 2008) provides the facility for reading a given SBML element's annotation and hence could be used to determine that `species1` and `species2` are annotated with ChEBI term CHEBI:4167 and KEGG Compound C00031, respectively. From this, it may be concluded that the compounds represented by these species are different. However, manual inspection of the database references in ChEBI and KEGG show that both species are annotated with references that share the same chemical structure, and hence are equivalent.

Performing such a comparison computationally is beyond the scope of libSBML. To do so, the annotations must be 'dereferenced' by querying the two databases via their web service interfaces. This task is complicated particularly because each of the web services has non-standard interfaces.

The libAnnotationSBML library creates a unified framework for supporting MIRIAM-compliant annotations by wrapping these divergent web services into a Java API, allowing each web service to be queried in a consistent manner. The library itself can act as a layer on top of the libSBML API.

The library is built dynamically by querying the MIRIAM web service (Laibe and Le Novère, 2007), which provides a collection of data types that are recommended for use in model annotation. The web service provides details of each of these data types

*To whom correspondence should be addressed.

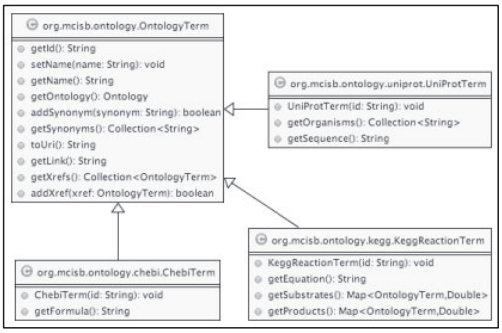

Fig. 2. Class diagram showing public methods of `OntologyTerm` and specialized subclasses `ChebiTerm`, `UniProtTerm` and `KeggReactionTerm`.

including names, URNs and physical URLs to resources. From this, a collection of `Ontology` objects are instantiated, one for each data type supported specified in MIRIAM.

Individual `OntologyTerms` objects are built up from an `Ontology` object and a unique identifier. Once instantiated, the `OntologyTerm` provides a number of methods, specified in Figure 2. The implementation of these methods is performed by mapping the calls to an appropriate call to the data type's web service, where such a web service exists.

The `OntologyTerm` class can be extended to provide methods specific to the SBML element that is being described. For example, a metabolite species element annotated with a ChEBI term will return a `ChebiTerm` object, providing a method for accessing the chemical formula of the metabolite. Similarly, a KEGG Reaction annotation will return a `KeggReactionTerm` object, providing methods for accessing reactants and products, each returned as `OntologyTerms` themselves.

Applying libAnnotationSBML to the SBML in Figure 1 will associate an `OntologyTerm` with each of the species. Calling `getName()` on these ChEBI and KEGG `OntologyTerm` objects returns 'D-glucopyranose' and 'D-glucose', respectively. Considering the initial problem of comparing SBML components, this provides an example of why names cannot be used reliably to perform this task. A more reliable approach is to exploit the fact that many data resources cross-reference one another. For example, entries in the ChEBI database can provide details of the equivalent term in KEGG, and vice versa. The `OntologyTerm` class supports this by implementing a `getXrefs()` method which returns cross references themselves as `OntologyTerms`, along with a predicate, defined in libSBML, that indicates the relationship between them. When an `OntologyTerm` references an equivalent entity in a different database, the predicate `libsbmlConstants.BQB_IS` is returned. In the case of a genomic database entry cross referencing an entry in a proteomic database, `libsbmlConstants.BQB_ENCODES` is used. Utilizing this method, it can be determined computationally that the ChEBI and KEGG terms cross-reference one another, and hence `species1` and `species2` can be unambiguously determined to represent equivalent entities.

The libAnnotationSBML library facilitates the rapid development of tools to manipulate SBML annotation terms. The library can be used to add annotation to unannotated SBML models, using a similar approach to semanticSBML (Schulz *et al.*, 2006). libAnnotationSBML can annotate both metabolites and proteins, exploiting the search facility that exists in both the ChEBI and UniProt web services (The UniProt Consortium, 2008).

The focus of libAnnotationSBML is to develop tools to manipulate already annotated models. An example of such a tool is the SbmlSynonymExtractor, which takes annotated SBML as input, and returns a mapping of all species terms to their name synonyms, harvested from ChEBI, KEGG or UniProt. Another tool, the SbmlReactionBalancer, determines whether the reactions specified within an SBML file are elementally balanced by querying the ChEBI web service to retrieve chemical formulae of reaction participants.

libAnnotationSBML was used extensively in the development of a genome-scale model of yeast metabolism, the first model of this scale in which all compartments, metabolites, enzymes and complexes are unambiguously defined using MIRIAM-compliant annotations (Herrgård *et al.*, 2008).

3 IMPLEMENTATION AND DISTRIBUTION

The API is written in Java 1.5 and is dependent upon libSBML v3. It is supported in Linux, Windows and MacOS X and is distributed as source code and associated build files under the open source Academic Free Licence v3.0 from http://mcisb.sf.net/ along with other tools described in this manuscript.

ACKNOWLEDGEMENTS

The authors thank the BBSRC and EPSRC for financial support of the Manchester Centre for Integrative Systems Biology, of which this work was an integral part.

Conflict of Interest: none declared.

REFERENCES

Bornstein,B.J. *et al.* (2008) LibSBML: an API Library for SBML. *Bioinformatics*, **24**, 880–881.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Herrgård,M. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnol.*, **26**, 1155–1160.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kanehisa,M. *et al.* (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kell,D.B. and Mendes,P. (2008) The markup is the model: reasoning about systems biology models in the Semantic Web era. *J. Theor. Biol.*, **252**, 538–543.

Laibe,C. and Le Novère,N. (2007) MIRIAM resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.*, **1**, 58.

Le Novère,N. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol.*, **23**, 1509–1515.

Schulz,M. *et al.* (2006) SBMLmerge, a system for combining biochemical network models. *Genome Inform. Ser.*, **17**, 62–71.

The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Publication 4

The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks.

Swainston N, Smallbone K, Mendes P, Kell DB, Paton NW.

J Integr Bioinform. 2011, **8**, 186.

The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks

Neil Swainston[1*], Kieran Smallbone[1], Pedro Mendes[1,2], Douglas B Kell[1], Norman W Paton[1]

[1] Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, M1 7DN, United Kingdom

[2] Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, VA 24061, USA

Summary

The generation and use of metabolic network reconstructions has increased over recent years. The development of such reconstructions has typically involved a time-consuming, manual process. Recent work has shown that steps undertaken in reconstructing such metabolic networks are amenable to automation.

The SuBliMinaL Toolbox (http://www.mcisb.org/subliminal/) facilitates the reconstruction process by providing a number of independent modules to perform common tasks, such as generating draft reconstructions, determining metabolite protonation state, mass and charge balancing reactions, suggesting intracellular compartmentalisation, adding transport reactions and a biomass function, and formatting the reconstruction to be used in third-party analysis packages. The individual modules manipulate reconstructions encoded in Systems Biology Markup Language (SBML), and can be chained to generate a reconstruction pipeline, or used individually during a manual curation process.

This work describes the individual modules themselves, and a study in which the modules were used to develop a metabolic reconstruction of *Saccharomyces cerevisiae* from the existing data resources KEGG and MetaCyc. The automatically generated reconstruction is analysed for blocked reactions, and suggestions for future improvements to the toolbox are discussed.

1 Introduction

The development of metabolic network reconstructions has increased over the last ten years. Such reconstructions are now available for a range of taxonomically diverse organisms, and they have been applied to a number of research topics including metabolic engineering, genome-annotation, evolutionary studies, network analysis, and interpretation of omics datasets [1].

A genome-scale metabolic reconstruction is a computational and mathematical model of the metabolic capabilities of a given organism [2]. It consists of all known metabolic reactions that can take place in a cell and the gene-protein-reaction relationships that connect the genome to the metabolome via the specification of enzymes and isoenzymes that catalyse each reaction. Specifying such gene-protein-reaction relationships will allow metabolic modelling to become increasingly integrated with transcription and signalling networks

[*] To whom correspondence should be addressed. Email: neil.swainston@manchester.ac.uk

through consideration of the action of metabolites on promoters and transcription factors. In addition, intra- and extra-cellular compartments can be considered, along with transport reactions and transport proteins that provide for metabolic transport across compartmental membranes. Furthermore, in order to analyse the phenotypic behaviour of the organism under a given condition, it is common to specify an objective function that is assumed to be optimised by the cell [3]. This can take a number of forms, including the maximisation or minimisation of usage of ATP, but commonly assumes that a cell attempts to maximise growth rate. In this case, a biomass function is included, which is a hypothetical reaction that uses metabolites necessary for cell growth, such as amino acids, nucleotides, lipids and cell-wall components, and required cofactors.

This work concerns itself with the automation of steps that are necessary in the development and analysis of genome-scale metabolic models. The process of completing such steps to develop reconstructions is now well defined and is recognised as being time-consuming [4]. While many of the steps associated with generating a high-quality reconstruction require manual curation, some of these are amenable to automation, providing the possibility of automating the process of generating a draft reconstruction to be used in subsequent manual curation. While a fully automated approach has been shown itself capable of the rapid generation of candidate reconstructions in a number of cases [5], it is recognised that such reconstructions still require manual validation and editing. As such, there remains a middle ground between the fully automated and fully manual approaches, where the draft reconstruction and curation process stands to benefit from dedicated software support. Such a semi-automated approach was followed in the development of recent genome-scale metabolic reconstructions for *Saccharomyces cerevisiae* and *Homo sapiens*, in which draft reconstructions were checked and enhanced by utilising SuBliMinaL Toolbox modules during an iterative development process.

Drawing upon previous experience of generating such reconstructions [6,7,8], this paper considers the development of reconstructions from the existing curated data resources KEGG [9] and MetaCyc [10]. Although both resources provide the facility for exporting metabolic models, neither of these exported models is of sufficient accuracy nor is suitably formatted for performing genome-scale, constraint-based analyses. Nevertheless, both resources provide initial pre-draft prototypes that can be developed further [11].

The SuBliMinaL Toolbox consists of a number of independent modules that can be used independently or chained together to form a reconstruction workflow allowing the generation of an initial draft of a metabolic reconstruction (see Figure 1). The importance of using community-developed standards to represent models in systems biology is well established [12]. As such, reconstructions are generated in Systems Biology Markup Language (SBML) [13] and are semantically annotated according to the MIRIAM standard [14]. They can be formatted in such a way that they can be loaded into the COBRA Toolbox [15], allowing constraint based analyses to be performed on the model, using techniques such as Flux Balance Analysis (FBA) [16].

2 Methods

SuBliMinaL Toolbox modules typically have a simple SBML-in / SBML-out interface, which take in a model or models, perform a given task and produce an updated model. Some modules should be used sequentially (for example, a reaction should only be elementally and charge balanced once the protonation states of its reactants and products have been determined). The SuBliMinaL Toolbox utilises the programming library libAnnotationSBML

[17], and web service interfaces to ChEBI [18] and KEGG to automatically retrieve required chemical data.

The SuBliMinaL Toolbox is written in Java, and is dependent upon third-party tools, which must be installed independently. Each of the modules can either be run from the command line or incorporated into custom software via a Java API. The SuBliMinaL Toolbox has been tested on Mac OS X 10.6 and 64-bit Windows 7. Instructions on the installation and use of the toolbox are available at http://www.mcisb.org/subliminal/. A description of each module of the toolbox is given below.

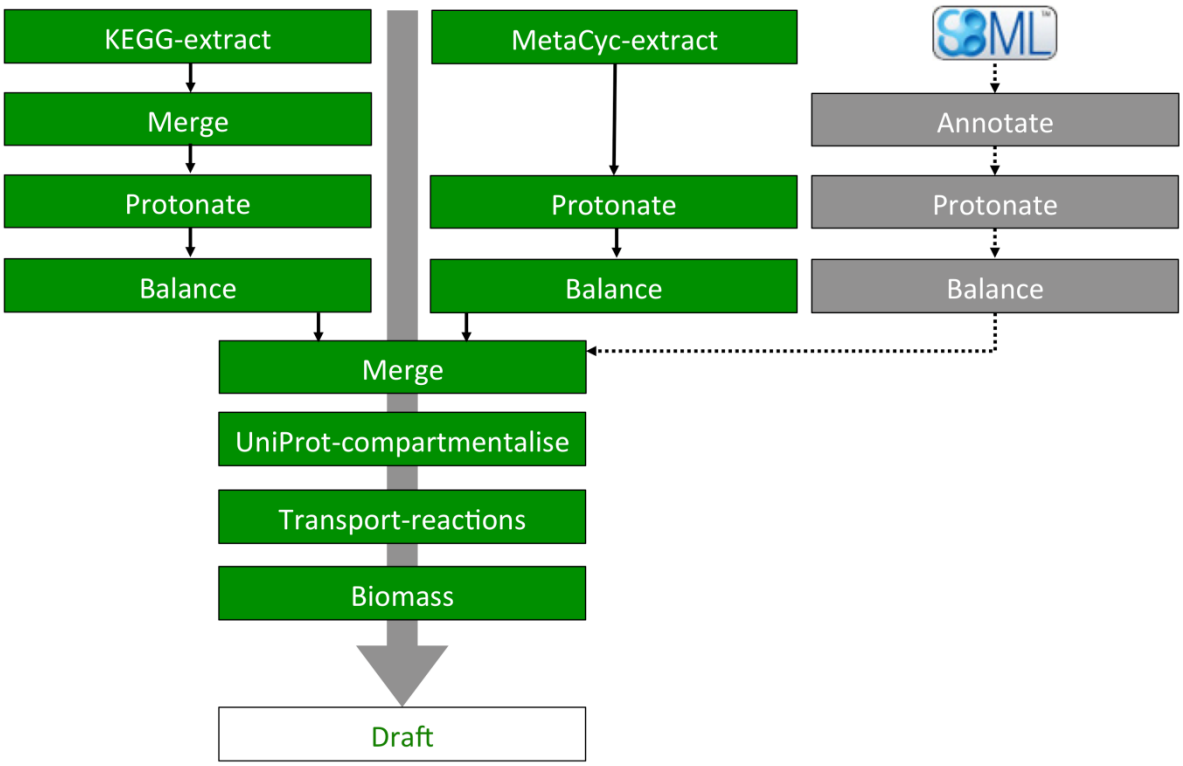

Figure 1: Flow diagram illustrating how SuBliMinaL Toolbox modules may be chained together to generate a draft metabolic network reconstruction that can be analysed in the COBRA Toolbox. The names of the boxes refer to individual SuBliMinaL Toolbox modules. The main branches with solid arrows from KEGG-extract and MetaCyc-extract indicate the pipeline that was utilised in this study. The right-hand branch with dotted arrows indicates a hypothetical addition to the pipeline, which could be used to include existing reconstructions or individual pathways marked up in SBML format.

2.1 Pre-draft reconstruction

Initial pre-draft pathways for a given organism can be generated from both KEGG and MetaCyc, using the **KEGG-extract** and **MetaCyc-extract** modules respectively.

KEGG does not allow export of pathways data in SBML format. The **KEGG-extract** module has been developed to provide this functionality. The module downloads the organism-specific KEGG KGML flat files for each represented pathway, and parses these to extract the individual metabolic reactions, in terms of metabolites and enzymes, that constitute the pathway. Where specified, reaction directionality is also considered. KEGG does not specify intracellular compartmentalisation, and as such, all metabolites are assumed to be cytoplasmic. An SBML model is then generated for each defined pathway, and each of these

is then annotated according to the MIRIAM standard, such that each metabolite and enzyme is assigned an unambiguous identifier.

It was found that both the existing tools for converting KEGG data into SBML format, KEGG2SBML (http://sbml.org/Software/KEGG2SBML) and KEGGConverter [19], were unsuitable for use in the context of generating genome-scale reconstructions, due to the presence of missing reactions or of reaction participants in their generated pathways. These shortcomings result in gaps and stoichiometric inconsistencies in the final draft reconstruction. Furthermore, both tools are reliant upon the downloading of KEGG flat files that are no longer freely available to academic users. Along with this work, the recently introduced KEGGtranslator [20] overcomes these limitations.

MetaCyc-extract downloads the appropriate organism-specific flat files and annotates the supplied SBML file to ensure consistency with the equivalent KEGG model. Again, the resulting model is updated to ensure appropriate metabolite charge state and balanced reactions. An advantage of MetaCyc over KEGG is in its definition of intracellular compartmentalisation. Where present, this intracellular compartmentalisation is extracted and added to the model. If specified, unambiguous metabolite and enzyme identifiers are extracted from the MetaCyc flat files and assigned to chemical species in the generated SBML file. If no identifiers are present, metabolite names are automatically searched against the ChEBI database in order to determine ChEBI identifiers to be assigned to metabolites. Checking against supplied chemical formulae validates the assignment of such identifiers.

Both modules generate consistently formatted models representing the union of all metabolic pathways described in each resource. These individual models can then be merged and their annotations exploited in subsequent modules.

2.2 Annotation

The modules of the toolbox are dependent upon the initial draft reconstruction being annotated with unambiguous identifiers according to the MIRIAM standard. In order to support the use of existing reconstructions and pathways in the toolbox, the **Annotate** module has been developed to automate the process of adding annotations to existing models. The **Annotate** model launches the SuBliMinaL Annotator, a graphical wizard that facilitates the annotation process. The SuBliMinaL Annotator allows the user to select a model in SBML format, which is then parsed to extract names of the model components compartments, metabolites and enzymes. Each of these terms is then searched against the databases Gene Ontology (GO) [21], ChEBI and UniProt [22] respectively. The results of these searches are presented to the user, allowing the selection of the appropriate database term with which to annotate the model component (see Figure 2). Upon completion of the annotation process, the updated model is saved and can be used with subsequent SuBliMinaL modules.

The **Annotate** module can also be run in "Silent mode", which does not rely on user selection of search results. In this mode, the name of the SBML element being searched is compared alphanumerically against the search results in a case-insensitive manner. Upon matching, the SBML element is annotated with the matching search term, allowing commonly named metabolites such as ATP to be quickly annotated without relying on user selection.

By default, the **Annotate** module assumes all chemical species to be metabolic. To specify that a chemical species represents a protein, the appropriate species elements in the SBML model must be annotated with the Systems Biology Ontology (SBO) Term `SBO:0000252` (denoting *polypeptide chain*) [23]. Doing so ensures that this species name is searched against UniProt. If the SBML model is annotated with an NCBI Taxonomy term [24], the UniProt

search is restricted to proteins of that organism. Figure 3 illustrates an SBML model that allows organism-specific searches of UniProt to be performed.

2.3 Model merging

The lack of consistent naming between components within existing reconstructions has been reported as an impediment to both manual and automatic comparison and construction of models [25], and was a major motivation for the use of semantic annotations that overcome this. As each of the initial pre-draft models generated by both **KEGG-extract** and **MetaCyc-extract** contain comparable identifiers, this issue is mitigated. The **Merge** module automatically merges each supplied model into a single consolidated model, in which duplicate metabolites, enzymes, and reactions are removed. If present, intracellular compartmentalisation is also considered, such that chemical species (metabolites and enzymes) are only considered to be duplicates if they share both identifier and compartment. As chemical species in different compartments are considered to be distinct, transport reactions are retained during the merge process.

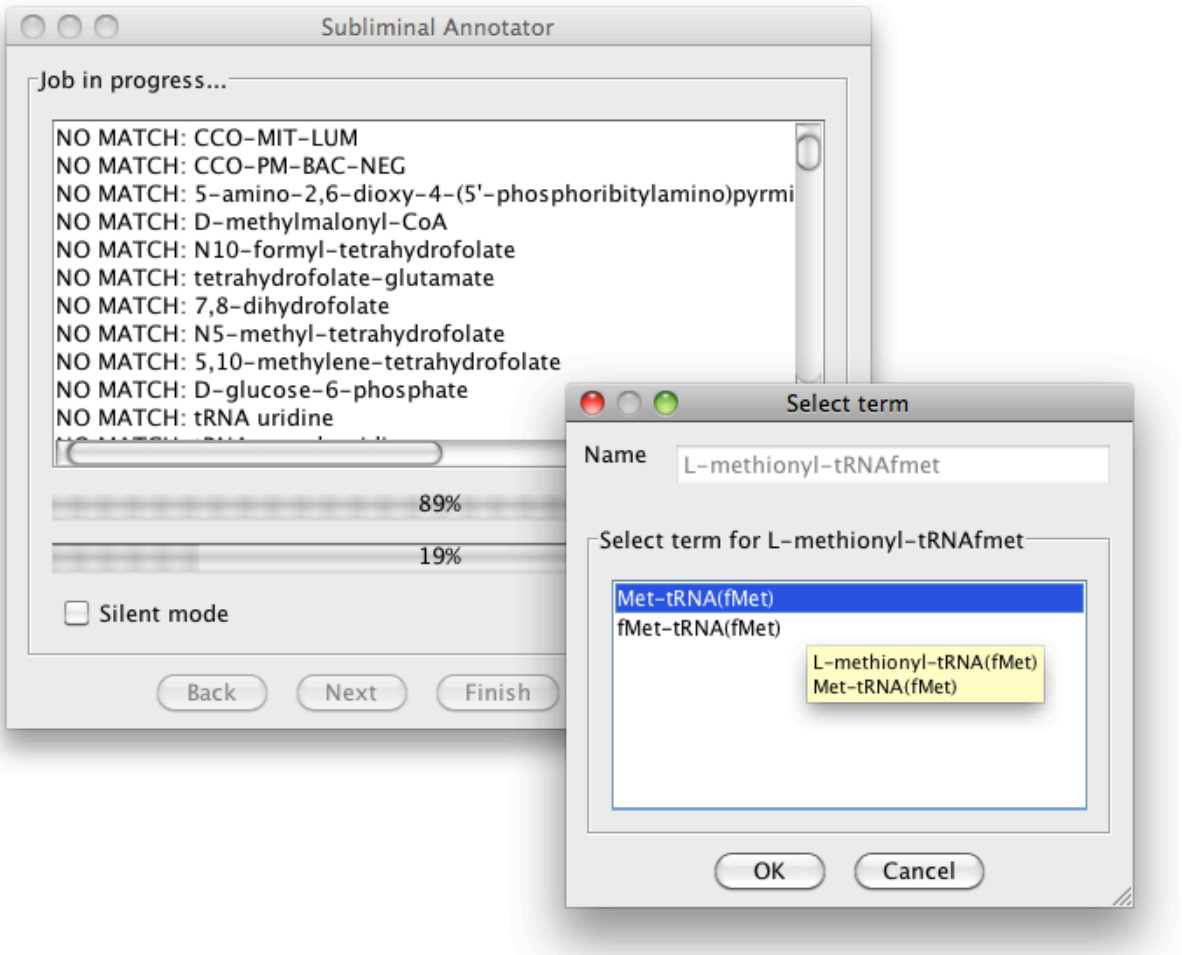

Figure 2: Screen capture of the SuBliMinaL Annotator. The main window displays progress of the annotation process, indicating terms that remain unmatched, and two progress bars displaying the percentage of terms successfully annotated and progress of the annotation process respectively. The foreground dialog box displays the results of a search for a metabolite name against ChEBI, ordered by the ChEBI Text Based Score. The user can select one of the two terms in order to annotate the metabolite. Compound synonyms are also searched, and can be viewed in a tooltip.

The merging of models is non-trivial due to the presence of duplicate metabolites both across data resources and within a given resource [26]. Furthermore, the specification of metabolites can differ in their precision. For example, in the case of KEGG, many stereoisomers are represented, an example being D-glucose, alpha-D-glucose and beta-D-glucose. The **Merge** module can therefore be run in a "fuzzy" mode, in which these metabolites are considered to be synonymous, as it is difficult to determine whether a given reaction involving these metabolites is intended to refer to the general case or one or both of the stereospecific terms.

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level2/version4" level="2" version="4">
  <model metaid="_model">
    <!-- The following specifies that the model represents Saccharomyces cerevisiae -->
    <annotation>
      <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
        <rdf:Description rdf:about="#_model">
          <bqbiol:is>
            <rdf:Bag>
              <rdf:li rdf:resource="urn:miriam:taxonomy:559292"/>
            </rdf:Bag>
          </bqbiol:is>
        </rdf:Description>
      </rdf:RDF>
    </annotation>
    <listOfCompartments>
      <compartment id="c" name="cytosol" size="1"/>
    </listOfCompartments>
    <listOfSpecies>

      <!-- The following species represent metabolites (small molecules) -->
      <species id="s1" name="D-Glucose 1-phosphate" compartment="c"/>
      <species id="s2" name="D-Glucose 6-phosphate" compartment="c" sboTerm="SBO:0000247"/>

      <!-- The following species represents a protein -->
      <species id="s3" name="Phosphoglucomutase-1" compartment="c" sboTerm="SBO:0000252"/>

    </listOfSpecies>
  </model>
</sbml>
```

Figure 3: Simple SBML model indicating how metabolites and enzymes are distinguished in the Annotate module by use of `sboTerm` attributes on the `species` elements. By default, all `species` are considered to be metabolic (small molecules) but this can be made explicit through use of the SBO term `SBO:0000247` (*simple chemical*). Enzymes are specified with the SBO term `SBO:0000252` (*polypeptide sequence*). Annotating the `model` element with an NCBI Taxonomy term (in this case, `559292`, representing *Saccharomyces cerevisiae*) limits the subsequent UniProt search to proteins belonging to the specified organism.

As such, the assumption can be made that a reaction applies to all synonymous metabolites, and these are then collapsed into a single metabolite in the merge process, with the intention of increasing the network connectivity of the merged reconstruction. The **Merge** module can determine whether two metabolites share the same chemical formula, and whether both have a shared ancestor in the ChEBI ontological tree. If so, these metabolites can be collapsed into a single term. For well-curated reconstructions, the **Merge** module can also be run in a more simplified mode, in which two metabolites will only be considered to be the same if they share the same semantic annotation.

2.4 Metabolite pKa prediction and determination of appropriate charge state

The **Protonate** module utilises the ChEBI web service to harvest SMILES strings [27] representing each metabolite. These are then passed to the `MajorMicrospeciesPlugin` method in the API of the cheminformatic library Marvin Beans for Java Developers (ChemAxon Kft., Budapest, Hungary; http://www.chemaxon.com), which relies on the Hammett−Taft approach [28] to estimate pKas and thus predict the dominant protonation state of the metabolites at a supplied pH. Specific pHs may also be applied to metabolites in a given intracellular compartment. Updated chemical formula and charge are then added to each metabolite, and, if appropriate, both the name and the ChEBI annotation of the molecule are updated to reflect its corrected charge state. Exploiting the ChEBI ontology specification of `isConjugateBaseOf` and `isConjugateAcidOf` predicates, which allow relationships between de/protonated molecules to be automatically determined, enables this functionality.

An example of this is KEGG compound `C00022`. Although KEGG names this metabolite pyruvate, both the molecular formula ($C_3H_4O_3$) and the cross-reference link to ChEBI (`CHEBI:32816`) for this entry indicate that the metabolite is actually the protonated form, pyruvic acid. Marvin Beans predicts that the metabolite is deprotonated at a pH of 7.0. As such, the **Protonate** module updates the metabolite in the reconstruction, setting the molecular formula to $C_3H_3O_3$, the charge to -1, and updates the annotation to that of the ChEBI term for pyruvate, `CHEBI:15361`. This illustrates the inconsistencies that are often present in biochemical resources, such that both conjugate acid and bases are sometimes collapsed into a single, ambiguous entry. Such inconsistencies can be resolved with this approach, producing unambiguous definitions of both metabolites and reactions that more accurately reflect physiological conditions.

2.5 Elemental and charge balancing

Balancing all metabolic reactions ensures that a reconstruction is free of stoichiometric inconsistencies [29]. Stoichiometric inconsistencies violate mass conservation, and can be illustrated in the example below:

$$R1: A \Leftrightarrow B$$

$$R2: A \Leftrightarrow B + C$$

It is intuitively clear that a network containing these two reactions contains an inconsistency. That is, metabolite C could only satisfy the above two equations if it were to have a mass and charge of zero. While the above example is simple, determining such errors in genome-scale models is non-trivial but can be performed algorithmically by the ScrumPy package [30]. While the ScrumPy package can detect such inconsistencies, their correction in the reconstruction relies upon manual curation. As such, it is preferable to reduce such inconsistencies by performing elemental and charge balancing where possible.

The **Balance** module automates this process of mass and charge balancing of reactions. Consider the following reaction:

$$\text{2-Acetolactate} + \text{carbon dioxide} \Leftrightarrow \text{Pyruvate}$$

$$C_5H_7O_4^- + CO_2 \Leftrightarrow C_3H_3O_3^-$$

Manual inspection can quickly determine that, in terms of elemental and charge balancing, two pyruvates must be produced, and the list of products is also deficient in a proton.

The **Balance** module attempts to detect (and fix) such issues automatically through mixed integer linear programming (MILP). Each reaction is represented as a matrix, A, containing elemental counts and charges for each reactant and product. Metabolites that are commonly absent from reaction definitions [31], such as water, protons and carbon dioxide, are also considered, and are added as both potential reactants and products. Reactant elemental and charge counts are specified as positive, those of products negative. Optional cofactor metabolites are only added to the matrix if they are not present in the original reaction. It is for this reason that carbon dioxide is absent from the specified optional reactants (see Figure 4).

The stoichiometric coefficients of each reactant are represented by the vector, b. Mixed integer linear programming is applied to solve $Ab = 0$, satisfying the constraint $b_j >= b_{j,min}$, where $b_{j,min}$ represents the minimum allowed stoichiometric coefficient for a given metabolite (1 for specified metabolites, 0 for optionally considered metabolites). This produces the vector of stoichiometric coefficients, b, to be applied to each reactant and product to balance the equation. (The vector, b, is minimised to return the minimum collection of stoichiometric coefficients that are required to balance the equation, thus preventing mathematically correct but non-optimal solutions, such as the spurious addition of water to both sides of the equation).

	Reactants		Products	Optional reactants		Optional products		
	CO2	C5H7O4	C3H3O3	H+	H20	H+	H20	CO2
C	**1**	**5**	**-3**	*0*	*0*	*0*	*0*	*-1*
O	**2**	**4**	**-3**	*0*	*1*	*0*	*-1*	*-2*
H	**0**	**7**	**-3**	*1*	*2*	*-1*	*-2*	*0*
charge	**0**	**-1**	**1**	*1*	*0*	*-1*	*0*	*0*
b_{min}^T	**1**	**1**	**1**	*0*	*0*	*0*	*0*	*0*

Figure 4: A matrix representing elemental count and charge of reactants and products in the reaction 2-Acetolactate + carbon dioxide ⇔ Pyruvate. Required reactant and product elemental and charge counts are specified in bold; those of optional reactants and products in italics. The vector b_{min}, representing the minimum permitted stoichiometric coefficients for each reaction participant, is also shown.

Consequently, the linear solver returns the solution, $b^T = (1\ 1\ 2\ 0\ 0\ 1\ 0\ 0)$, indicating that in order to balance the reaction, two pyruvates should be produced and one proton should be added as a product. In the case of a solution being found, the reaction is updated in the reconstruction to reflect this:

2-Acetolactate + carbon dioxide ⇔ 2 Pyruvate + proton

$$C_5H_7O_4^- + CO_2 \Leftrightarrow 2\ C_3H_3O_3^- + H^+$$

In many cases, however, reactions cannot be balanced with the above approach. This could be due to a number of reasons. An obvious limitation occurs when attempting to balance reactions in which the chemical formula of one or more participants is unknown, which is a result of missing information in the data resources. A further trivial problem is the specification of incorrect reactions, in which key reactants or products, over and above commonly absent metabolites such as water, are missing. In both cases, manual curation will be necessary to correct the errors, and calculating an elemental difference between the

reactants and products, which could suggest the chemical formula of a missing participant, may drive this process.

The **Balance** module uses the linear solver glpk (http://www.gnu.org/s/glpk/) and the java interface GLPK for Java (http://glpk-java.sourceforge.net).

2.6 Compartmentalisation

Thus far, the SuBliMinaL Toolbox generates largely uncompartmentalised reconstructions. Some intracellular compartmentalisation is provided by MetaCyc, but given the dependency of the pipeline described in figure 1 on KEGG, which does not consider compartmentalisation, most metabolites are considered to be cytoplasmic by default.

The Compartmentalise modules provide the facility for extending reconstructions generated from KEGG or MetaCyc alone to generate semi-compartmentalised models. Two compartmentalise modules exist. **UniProt-compartmentalise** extracts protein localisation information directly from UniProt annotation. Only Swiss-Prot entries are considered, and as the metadata associated to these entries is manually curated, the localisation specified for such entries is therefore likely to be accurate. For cases where no such curated data exist, the **PSORT-compartmentalise** module can be used. This module harvests protein sequences from the UniProt web services for each enzyme and passes these to the protein localisation service WoLFPSORT [32], a web interface to the PSORT algorithm [33]. From the curated or predicted intracellular compartmentalisation of a given enzyme, the localisation of metabolic reactions catalysed by this enzyme is inferred. As such, reactions, enzymes and metabolites are localised, and can be inferred to be present in multiple compartments, depending on the UniProt annotation or prediction of WoLFPSORT. In the case of a reaction being catalyzed by isoenyzmes that are present in different intracellular compartments, the reaction is duplicated such that an instance appears in each compartment, with the appropriate isoenzyme specified as the reaction modifier. Where predictions suggest that metabolites are found in multiple compartments, putative intracellular transport reactions are added to the reconstruction to allow for their transport between compartments.

2.7 Transport

Transport reactions are important for both natural metabolites and xenobiotics [34,35]. The **Transport-reaction** module adds a generic set of import reactions to the reconstruction in order to allow for uptake of metabolites from the growth medium. The set of generic import reactions are taken from the BIGG database [36], which contains 9 published and well-curated reconstructions from a range of taxonomically diverse organisms[1]. Import reactions across the cell membrane are added if the extracellular metabolite is also present in the reconstruction's cytoplasm. The addition of this generic set of import reactions is essential if subsequent analysis by the COBRA Toolbox is to be performed, as neither KEGG nor MetaCyc provide such cell-membrane transport reactions, which effectively means that reconstructions generated from these resources would be "starved" of growth media metabolites.

[1]These reconstructions are *S. cerevisiae* iND750, *E. coli* iAF1260, *E. coli* iJR904, *E. coli* textbook, *H. pylori* iIT341, *H. sapiens* Recon_1, *M. barkeri* iAF692, *M. tuberculosis* iNJ661, and *S. aureus* iSB619. Uptake reactions specific to the *H. sapiens* reconstruction were excluded from the set of selected uptake reactions, as they accounted for a number of metabolites for which transporters would be unlikely to be present in the majority of organisms.

Irreversible export reactions are added for all cytoplasmic metabolites, providing by default the possibility of excreting all cytoplasmic metabolites from the cell. This approach relies upon no *a priori* knowledge of the transport capabilities of the cell, and follows the philosophy of Fell *et al.* [37], which states that a simple solution to the problem of adding reactions to a reconstruction is to "add more than is likely to be necessary and to remove at a later date the ones that are not functional". It is envisaged that subsequent flux balance analysis of the completed draft model will provide an indication of which intracellular metabolites will need to be excreted in order for the model to fulfill the objective function. Superfluous export reactions can then be purged from the model, leaving a subset that can be manually validated according to the known capabilities of the organism, which may have been tested experimentally by metabolic footprinting [38]. The approach of adding more transport reactions than may be biologically feasible mirrors that of the **compartmentalise** modules, in which compartments are added with the intention of removing or reconciling these later as the reconstruction is manually validated.

2.8 Biomass function

The **Biomass** module assigns a generic biomass reaction to the reconstruction, and performs reformatting that allows FBA simulations to be performed by the COBRA Toolbox. The generic biomass reaction consists of the 20 most common amino acids, the four nucleotide precursors of DNA, ATP and lipids. In addition, the biomass reaction contains ADP, phosphate and protons as products. These "by-products" of biomass formation are then subsequently available to the model.

While the first biomass components are static and are applied to all reconstructions, the lipid term is built dynamically, and is dependent upon the constituents of the reconstruction. The generic metabolite "lipid" is first added to the list of biomass components. A number of modelling reactions are then added to the reconstruction, in which any metabolites in the model that exhibit an "is a" lipid relationship in the ChEBI database are added as reactants, with lipid as product.

Each of the reactants and products in the biomass reaction are given a stoichiometry of 1. This simple approach allows the reconstruction to be analysed to determine network connectivity, i.e., testing if the reconstruction allows for growth of the organism under given conditions. However, by not quantifying the components in biomass relative to one another, the reconstruction is unable to predict growth rate. This limitation can be rectified by manual curation.

3 Results

From applying the pipeline illustrated in Figure 1, a draft version of a metabolic reconstruction for *Saccharomyces cerevisiae* was generated for comparison against a manually generated version [7], which has been updated iteratively over a number of years. A comparison of both models is given in Table 1.

While SuBliMinaL generates a model with an increased number of metabolites and metabolic reactions with respect to the manually generated version (an increase of 92% and 90% respectively), it remains unclear whether this increase is due to the combined coverage of the original resources, KEGG and MetaCyc, or an indication of incomplete merging of data from each source. While the **Merge** module attempts to ensure that duplicate metabolites and reactions are not added to the consensus, metabolites that are lacking in comparable identifiers across the two sources may be duplicated. A limitation of both KEGG and

MetaCyc (and hence also of reconstructions generated from these resources) is the lack of defined multimeric enzymatic complexes.

While reactions are associated with genes and proteins where possible, specification of multimeric complexes in reconstructions developed by SuBliMinaL remains a task for manual curation, as it appears that no data resource describing such complexes currently exists, preventing the automation of this step.

Table 1: Comparison of SuBliMinaL- and manually-generated *S. cerevisiae* metabolic reconstructions. Values for unique metabolites, enzymes and metabolic reactions refer to "flattened" versions of the reconstructions, in which metabolites and enzymes in different intracellular compartments are considered one. In the case of SuBliMinaL, unblocked reactions were calculated on a minimal growth medium as described below.

Components	SuBliMinaL[2]	Manual
Compartments	8	17
Unique metabolites	1397	728
Unique enzymes	936	939
Unique metabolic reactions	1803	947
Unblocked reactions	1428/1803 (79%)	759/947 (80%)

The manually generated reconstruction also contains 9 compartments in addition to those in the SuBliMinaL-generated version. This is due to the specification of membrane compartments in the manual version, in which transport proteins are assigned. SuBliMinaL assumes all transport proteins to be present in the cytoplasm. This is simply a design decision to reduce the complexity of the reconstruction, and has no effect on its subsequent analysis.

A goal of the pipeline was to generate a reconstruction that was capable of simulating the production of biomass from minimal growth media automatically. It was found that the reconstruction could successfully simulate biomass production from a growth medium of D-glucose, ammonium, phosphate, sulphate, oxygen, water and protons. In doing so, it was found that, of the putative extracellular transport reactions added by the **Transport-reaction** module, all but 12 could be removed for the objective to be realised. The retention of these putative extracellular transport reactions provide sinks for product metabolites that are generated in reactions required to fulfill the biomass objective function. Of these 12 extracellular reactions that had to be retained, 3 involved metabolites involved in purine metabolism, suggesting reactions in this pathway that are incapable of carrying flux, which could act as a starting point for manual curation efforts.

The `fluxVariability` functionality of the COBRA Toolbox was used to assess the reconstruction. In order for a metabolic reaction to carry flux, all of its reactants and products must be connected to other reactions. As such, the proportion of reactions with capacity to carry flux is a measure of the connectivity of the network. The SuBliMinaL-generated reconstruction is found to be highly connected (75% unblocked), though slightly less than the manually curated version (80%).

[2]The SuBliMinaL-generated *S. cerevisiae* model was generated with KEGG release 59.0 (1 Jul 2011), MetaCyc version 15.1 (8 June 2011) and ChEBI release 83 (5 September 2011).

4 Discussion

The finding that the draft reconstruction contains suspected gaps in certain pathways illustrates the philosophy behind the development of draft reconstructions that are capable of undertaking constraint-based analysis: that is, that the results of such analyses can drive the curation process in an iterative manner through performance of cycles of analysis and refinement. The results of the analyses can be inspected, indicating potential errors, which can then be corrected manually.

The feasibility of performing such iterative cycles is made possible due to the speed at which genome-scale reconstructions can be automatically generated and checked. The pipeline described in Figure 1 generated the draft yeast reconstruction in under four hours on an Apple MacBook Pro 2.8GHz Intel Core i7. An existing protocol describing the generation of metabolic reconstructions suggests that the manual reconstruction refinement stage can take months to a year to complete [39]. This stage contains a number of steps that are covered by the SuBliMinaL Toolbox (such as charge state determination, reaction balancing, determination of metabolic identifiers), and as such, use of the toolbox should reduce the duration of both the initial stage of generating draft reconstructions and the checking of reconstructions in the following refinement phase.

The SuBliMinaL Toolbox has been used in the development of community-produced reconstructions of *Saccharomyces cerevisiae* and *Homo sapiens*. The use of the toolbox could be extended to the incremental development of such resources. As the development of reconstructions is an ongoing process, one could imagine a paradigm in which domain experts extract individual pathways from reconstructions, and then perform validation and curation on those areas of metabolism in which they have expertise. Such curated pathways could subsequently be re-collated into the reconstruction, which would then be formatted and reanalysed, following the iterative cycle described above. Such crowd-sourcing methods are already exploited in the web-based curation of individual pathways [40,41] and would prove useful in the iterative development of larger-scale networks. The recently developed software MEMOSys [42] may support such an approach, providing a secure web-enabled interface for community driven, multi-user development and refinement of reconstructions. The use of such tools, combined with automated modules for performing such tasks as checking of metabolite charge state determination and reaction balancing described here, may mitigate the need for jamborees: focused curation meetings that have become the preferred method of developing community-driven reconstructions over recent years [43].

Nevertheless, before such a more automated approach to community development could become more prevalent, there remain a number of issues within modules of the SuBliMinaL Toolbox that will need to be improved over time. While some reaction directionality is specified by KEGG, other reactions are initially specified to be reversible, which can result in thermodynamically infeasible flux patterns being predicted by model analyses. Specifying reaction directionality – either by automated or manual means - adds constraints to the model, which are likely to improve the model's predictive accuracy by preventing reactions that are thermodynamically infeasible. Due to the accessibility of InChI strings for many of the metabolites present in the reconstructions generated by the toolbox, there exists the possibility of automating the determination of reaction directionality, following the approach of Fleming *et al.* [44,45]. Integration of predictions of reaction directionality is therefore a likely future development.

The **Compartmentalise** module provides a useful first pass at automating the generation of compartmentalised reconstructions. While such an approach is preferable to a fully manual

approach to determining compartmentalisation that is currently followed, it is recognised that this approach is completely dependent upon the accuracy of the UniProt annotation or WoLFPSORT predictions. It is therefore likely that such an approach would reduce the connectivity of the network, as many pathways in a given intracellular compartment would be incomplete unless all enzymes within the pathway were correctly predicted to be present in the compartment. Applying this approach would require the addition of missing reactions in a gap-filling step, which may be performed by inference. For example, if an almost complete mitochondrial TCA cycle was predicted to be missing an enzyme that is present in the cytoplasm, it may be inferred that the enzyme (and the reaction that it catalyses) is indeed present in the mitochondria, despite the UniProt annotation or WoLFPSORT algorithm predicting otherwise. Such inferred reactions would be added as "modeling" reactions, and associated with an appropriate evidence code, indicating that they should be subject to subsequent manual curation. The inference of enzyme and reaction localisation, based upon the network topology of partially compartmentalised metabolic models, has been reported [46], and the approach followed by the toolbox – to generate a partially compartmentalised model for subsequent refinement – supports this inference method.

Limitations of the current **Biomass** module implementation are its assumption that *all* lipids may be constituents of the biomass objective function, and that other cell wall constituents and storage carbohydrates are not considered. Determining more specific biomass objective functions, perhaps tailored towards the taxonomy of the organism under reconstruction, would be a useful improvement for future work.

The possibility exists to extend the toolbox to consider transport proteins. Transport proteins for a given organism can be automatically extracted from the TransportDB database [47] and the potential exists to add these to the reconstruction. However, while the transport proteins can be extracted, TransportDB does not yet fully characterise its transport proteins in such a way that the corresponding transported metabolites can be retrieved in an automated fashion. It is hoped that, as such resources that describe transport proteins develop, the task of assigning such proteins to individual reactions will be able to be automated.

It is again emphasised that manual curation and validation are essential steps in generating a high-quality reconstruction. Referring to literature commonly drives this validation process, and recently developed reconstructions have illustrated the importance of applying literature references and confidence scores to components within the model. Doing so increases the confidence that users apply to reconstructions (or at least, individual pathways or reactions within reconstructions), and also can be used to prioritise refinement efforts. The determination of literature references may be aided through tighter integration with text-mining tools such as PathText [48] in order to simplify the arduous, but necessary, task of finding evidence for present (and missing) reactions in the literature [49].

Acknowledgements

The authors thank the BBSRC and EPSRC for their funding of the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org), BBSRC/EPSRC Grant BB/C008219/1. The authors also thank Michael Howard and Daniel Jameson for help with the preparation of the manuscript.

References

[1] M. A. Oberhardt, B. Ø. Palsson, J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320, 2009.

[2] M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, B. Ø. Palsson. Metabolic modeling of microbial strains in silico. *Trends Biochem Sci*, 26:179-86, 2001.

[3] R. Schuetz, L. Kuepfer, U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol*, 3:119, 2007.

[4] I. Thiele, B.Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5:93-121, 2010.

[5] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, R. L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*, 28:977-982, 2010.

[6] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, D. B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*, 26:1155-1160, 2008.

[7] P. D. Dobson, K. Smallbone, D. Jameson, E. Simeonidis, K. Lanthaler, P. Pir, C. Lu, N. Swainston, W. B. Dunn, P. Fisher, D. Hull, M. Brown, O. Oshota, N. J. Stanford, D. B. Kell, R. D. King, S. G. Oliver, R. D. Stevens, P. Mendes. Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol*, 4:145, 2010.

[8] I. Thiele I, D. R. Hyduke, B. Steeb, G. Fankam, D. K. Allen, S. Bazzani, P. Charusanti, F. C. Chen, R. M. Fleming, C. A. Hsiung, S. C. De Keersmaecker, Y. C. Liao, K. Marchal, M. L. Mo, E. Özdemir, A. Raghunathan, J. L. Reed, S. I. Shin, S. Sigurbjörnsdóttir, J. Steinmann, S. Sudarsan, N. Swainston, I. M. Thijs, K. Zengler, B. O. Palsson, J. N. Adkins, D. Bumann. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst Biol*, 5:8, 2011.

[9] M. Kanehisa, S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28:27-30, 2000.

[10] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, S. M. Paley, A. Pellegrini-Toole. The EcoCyc and MetaCyc databases. *Nucleic Acids Res*, 28:56-59, 2000.

[11] K. Radrich, Y. Tsuruoka, P. Dobson, A. Gevorgyan, N. Swainston, G. Baart, J. M. Schwartz. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol*, 4:114, 2010.

[12] D. B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov Today*, 11:1085-92, 2006.

[13] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M.

Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang J; SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524-531, 2003.

[14] N. Le Novère, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence, B. L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol*, 23:1509-15, 2005.

[15] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B.Ø. Palsson, M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*, 2:727-38, 2007.

[16] J. D. Orth, I. Thiele, B. Ø. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28:245-8, 2010.

[17] N. Swainston, P. Mendes. libAnnotationSBML: a library for exploiting SBML annotations. *Bioinformatics*, 25:2292-2293, 2009.

[18] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36:D344-350, 2008.

[19] K. Moutselos, I. Kanaris, A. Chatziioannou, I. Maglogiannis, F. N. Kolisis. KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. *BMC Bioinformatics*, 10:324, 2009.

[20] C. Wrzodek, A. Dräger, A. Zell. KEGGtranslator: visualizing and translating the KEGG PATHWAY database. *Bioinformatics*, 27:2314-2315, 2011.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25-9, 2000.

[22] UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, 38:D142-148, 2010.

[23] N. Le Novère. Model storage, exchange and integration. *BMC Neuroscience*, 7:S11, 2006.

[24] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y . Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, J. Ye J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 37:D5-15, 2009.

[25] L. Kuepfer. Towards whole-body systems physiology. *Mol Syst Biol*, 6:409, 2010.

[26] M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel, D. A. Fell. Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEE Proc Syst Biol*, 153:379-84, 2006.

[27] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28:31-36, 1988.

[28] F. Csizmadia, A. Tsantili-Kaoulidou, I. Paderi, F. Darvas. Prediction of distribution coefficient from structure. 1. Estimation method. *J Pharm Sci*, 86:865– 871, 1997.

[29] A. Gevorgyan, M. G. Poolman, D. A. Fell. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 24:2245-2251, 2008.

[30] M. G. Poolman. ScrumPy: metabolic modelling with Python. *IEE Proc Syst Biol*, 153:375-378, 2006.

[31] M. A. Ott, G. Vriend. Correcting ligands, metabolites, and pathways. *BMC Bioinformatics*, 7:517, 2006.

[32] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, K. Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, 35:W585-7, 2007.

[33] K. Nakai, P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24:34-36, 1999.

[34] P. D. Dobson, D. B. Kell. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Discov*, 7:205-20, 2008.

[35] P. D. Dobson, K. Lanthaler, S. G. Oliver, D. B. Kell DB. Implications of the dominant role of transporters in drug uptake by cells. *Curr Top Med Chem*, 9:163-81, 2009.

[36] J. Schellenberger, J. O. Park, T. M. Conrad, B.Ø. Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11:213, 2010.

[37] D. A. Fell, M. G. Poolman, A. Gevorgyan. Building and analysing genome-scale metabolic models. *Biochem Soc Trans*, 38:1197-201, 2010.

[38] J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, D. B. Kell. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol*, 21:692-6, 2003.

[39] I. Thiele, B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5:93-121, 2010.

[40] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, C. Evelo. WikiPathways: pathway editing for the people. *PLoS Biol*, 6:e184, 2008.

[41] Y. Matsuoka, S. Ghosh, N. Kikuchi, H. Kitano. Payao: a community platform for SBML pathway model curation. *Bioinformatics*, 26:1381-3, 2010.

[42] S. Pabinger, R. Rader, R. Agren, J. Nielsen, Z. Trajanoski. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC Syst Biol*, 5:20, 2011.

[43] I. Thiele, B. Ø. Palsson. Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol*, 6:361, 2010.

[44] R. M. Fleming, I. Thiele, H. P. Nasheuer. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. *Biophys Chem*, 145:47-56, 2009.

[45] R. M. Fleming, I. Thiele. von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. *Bioinformatics*, 27:142-3, 2011.

[46] S. Mintz-Oron, A. Aharoni, E. Ruppin, T. Shlomi. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics*, 25:i247-52, 2009.

[47] Q. Ren, K. Chen, I. T. Paulsen. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res*, 35:D274-279, 2007.

[48] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, J. Tsujii. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26:i374-381, 2010.

[49] C. Nobata, P. D. Dobson, S. A. Iqbal, P. Mendes, J. Tsujii, D. B. Kell, S. Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, 7:94-101, 2001.

Publication 5

SBML Level 3 Package Proposal: Annotation.

Waltemath D*, **Swainston N***, Lister A*, Bergmann F, Henkel R, Hoops S, Hucka M, Juty N, Keating S, Knuepfer C, Krause F, Laibe C, Liebermeister W, Lloyd C, Misirli G, Schulz M, Taschuk M, Le Novère N.

Nature Precedings. 2011 http://dx.doi.org/10.1038/npre.2011.5610.1.

**Equal contribution.*

SBML Level 3 Package Proposal: Annotation

Dagmar Waltemath[1,*], Neil Swainston[2,*], Allyson Lister[3,*], Frank Bergmann[4], Ron Henkel[1], Stefan Hoops[5], Michael Hucka[6], Nick Juty[7], Sarah Keating[7], Christian Knuepfer[8], Falko Krause[9], Camille Laibe[7], Wolfram Liebermeister[9], Catherine Lloyd[10], Goksel Misirli[3], Marvin Schulz[9], Morgan Taschuk[3], Nicolas Le Novère[7]

[1]Database and Information Systems, University of Rostock, 18051 Rostock, MV, Germany

[2]Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, UK

[3]School of Computing Science, Newcastle University, Newcastle-Upon-Tyne, UK

[4]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

[5]Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, VA 24061, USA

[6]Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125, USA

[7]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[8]Institute of Computer Science, Friedrich-Schiller-University Jena, Germany

[9]Institut für Biologie, Theoretische Biophysik, Humboldt-Universität zu Berlin, Berlin, Germany

[10]Auckland Bioengineering Institute, University of Auckland, Auckland 1010, New Zealand

[*]These authors contributed equally to this work.

Proposal title

SBML Level 3 Annotation Package. (Keyword: annot).

Proposal tracking number

Number 3009839 in the SBML issue tracking system.

Version information

Version number and date of public release

This is version 1 of the Annotation package proposal. It reflects the results of the Annotation package meeting, 19–21 May 2010.

URL or this version of the proposal

Annot package proposal version 1 (2011-02).

URL for the previous version of this proposal

None.

Introduction and motivation

Annotations encode meta-information in SBML models. SBML allows users to annotate any SBML component that extends `SBase` (SBML L3 V1 Core specification, p. 15). The *Annotation* concept provides a container for optional software-generated, computer-readable content not meant to be shown to humans. The current syntax for encoding of information inside the annotation element, hereafter referred to as *Core annotation* recommends the use of a defined subset of RDF as described in the SBML L3 V1 Core specification, section 6. The Core annotation format allows the expression of relationships between SBML elements and resources referred to by values of `rdf:resource` attributes. The BioModels.net relation

qualifiers (predicates) (http://biomodels.net/qualifiers/) define the nature of the relationship (SBML L3 V1 Core specification, p. 87).

However, as annotations are independent from the model syntax and are not required for successful simulation of the models, it is proposed that it would be more suitable to define annotations in its own package. It is proposed to retain Core annotations in the SBML Level 3 Core, but to develop a Level 3 extension package to extend the possibilities of Core annotations and therefore support a richer set of meta-information that are currently not expressible. In future Levels, the original Core annotations may be completely replaced by this package.

Background

The package builds on the description of the Core annotation as currently described in the SBML L3 V1 Core specification, section 6. A short description of the Core annotation standard follows after the introduction to RDF.

Introduction to RDF

The Resource Description Framework (RDF) is a language for representing information about resources, in particular for representing metadata about web resources in the World Wide Web. The RDF Primer generalises the concept of a "web resource" to represent information about things that can be identified on the web, even when they cannot be directly retrieved on the web. RDF-encoded information can be processed by applications. The common framework provided by RDF to express the information in a standardised way leverages the loss-less exchange of information between different applications. RDF builds upon ideas from knowledge representation, artificial intelligence, and data management.

RDF Statements

The basic concept of RDF is the identification of things using Uniform Resource Identifiers (URIs). The resources are described by properties with particular property values. The specific terminology used in RDF is (see RDF Primer, section 2.1):

- **subject**: The part that identifies the thing the statement is about is called the subject.
- **predicate**: The part that identifies the property of the subject that the statement specifies is called the predicate.
- **object**: The part that identifies the value of a property is called the object.

Because of the generality characteristic of URIs, they are used in RDF to identify subjects, predicates and objects in statements. RDF statements effectively take the form of triples, allowing statements to be written in the form:

- **subject** has **predicate** whose value is **object**.

The RDF primer extends the concept of URIs to URI references, which are defined as:

- **`URIref`**: A URI reference (or `URIref`) is a URI, together with an optional fragment identifier at the end. The fragment is separated by the # character.

RDF URIs can be used to encode different kinds of information, including kinds of things, individuals, properties of things, or values of properties.

RDF refers to a resource as:

- **`resource`**: A `resource` is defined as anything that is identifiable by a URI reference (`URIref`).

Objects in RDF may either be `URIrefs`, or constant values (literals). Subject and predicate cannot both be literals. Using `URIrefs` as subject, predicate and object in statements supports the development and use of shared vocabularies on the web. One advantage of using `URIrefs` for statement definitions is that an `URIref` allows for the more precise identification of a thing than using a sole string (e.g. `http://www.ex.org/staffif/1111` identifying a person more precisely than the string "Eric Miller").

RDF notations

RDF allows the encoded information to be modelled in different ways. One way is the representation of the information as a graph of nodes and arcs. An RDF graph is formed based on the idea that *the things being described have properties which have values, and that resources can be described by making statements [...] that specify those properties and values* (RDF Primer, section 2.1). The nodes in the graph represent the **subject** and **object** of a statement. The arc represents the **predicate**. It is directed from **subject** node to **object** node. Ellipses in the RDF graph represent `URIrefs`, while boxes represent literals. A sample RDF graph is shown in Figure 1.

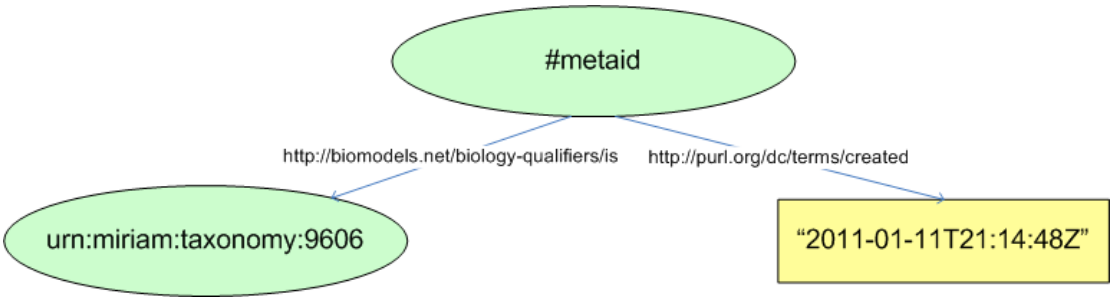

Figure 1: An example RDF graph utilising both a `URIref` and a literal.

A second way to represent RDF statements is the use of the triplet notation. It offers an alternative to the graph representation; e.g. if a graph gets too inconvenient to be drawn. Each statement of the graph is written as a single triple, consisting of the **subject**, **predicate** and **object** (in that order). A triple describes a single arc in the graph, with the **subject** being the arc's beginning and the **object** being the arc's ending. `URIrefs` are put in angle brackets (`<...>`), while literals are put in quotes (`"..."`). Examples of such notation, as RDF triples, are:

Subject	Predicate	Object
`<#metaid>`	`<http://biomodels.net/biology-qualifiers/is>`	`<urn:miriam:taxonomy:9606>`
`<#metaid>`	`<http://purl.org/dc/terms/create>`	`"2011-01-11T21:14:48Z"`

Furthermore, XML can be used to represent statements in a machine readable way. The syntax for writing RDF in XML is called RDF/XML (see RDF/XML Syntax Specification). The description of a statement is enclosed in an `rdf:RDF` XML element. The statement itself is enclosed in an `rdf:Description` element; being regarded a description about the **subject** of the statement. The **subject** is referred to in the `rdf:about` attribute inside the `rdf:Description` element. The property element representing the **predicate** and **object** of the statement is nested within the containing `rdf:Description` element. The nesting indicates the application of the property on the given subject. More details on the RDF/XML syntax are given in the RDF Syntax Specification.

An example of RDF/XML representation, marking up the two statements above, is:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
  <rdf:Description rdf:about="#metaid_recon1_1">
    <bqbiol:is rdf:resource="urn:miriam:taxonomy:9606"/>
    <dcterms:created>2011-01-11T21:14:48Z</dcterms:created>
  </rdf:Description>
</rdf:RDF>
```

Figure 2: An example of RDF/XML utilising both a `URIref` and a literal.

SBML Core annotation standard

According to the current SBML Core annotation standard, RDF/XML is used to present the RDF statements (see Figure 2, taken from the SBML L3 V1 Core Specification, p.86).

```
<SBML_ELEMENT +++ metaid="SBML_META_ID" +++ >
  +++
  <annotation>
    +++
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:dc="http://purl.org/dc/elements/1.1/"
             xmlns:dcterm="http://purl.org/dc/terms/"
             xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#"
             xmlns:bqbiol="http://biomodels.net/biology-qualifiers/"
             xmlns:bqmodel="http://biomodels.net/model-qualifiers/" >
      <rdf:Description rdf:about="#SBML_META_ID">
        [HISTORY]
        <RELATION_ELEMENT>
          <rdf:Bag>
            <rdf:li rdf:resource=" URI " />
            ...
          </rdf:Bag>
        </RELATION_ELEMENT>
        ...
      </rdf:Description>
      +++
    </rdf:RDF>
    +++
  </annotation>
  +++
</SBML_ELEMENT>
```

Figure 3: SBML L3 V1 Core annotation standard.

The current Core annotation schema, while written in RDF/XML, supports only a limited subset of RDF/XML. The above syntax must be followed, including the use of the mandatory `rdf:Bag` container, and the specification of the **subject** as a URI in the `rdf:li` `rdf:resource` attribute.

The URI link to an external resource must be perennial. To uniquely identify a controlled vocabulary term or object, the Minimum Information Required in the Annotation of Models (MIRIAM) standard is used[1]. A referenced MIRIAM URI maps to a physical web source, i.e. a URL. The connection between the addressed third-party knowledge and the annotated element is established using any of the model or biological qualifiers listed on http://www.biomodels.net/qualifiers/. If an annotation follows the proposed scheme, it is considered an SBML MIRIAM annotation. The SBML history element enables the tracking of changes as it allows the storage of the annotation creators and modification dates.

Problems with Core annotation

Statements about attributes

The Core annotation specification reuses the RDF approach of providing `rdf:Description` elements for SBML XML elements, such as species or compartment.

However, there currently does not exist a mechanism to annotate SBML attributes. See, for example, the following SBML code snippet:

```
<species metaid="metaid_0000042" id="Y" name="Intravesicular
```

[1] Le Novère N, 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 JL, Spence HD, Wanner BL. Minimum Information Requested In the Annotation of biochemical Models (MIRIAM). *Nature Biotechnology* 2005, **23**, 1509-1515.

```
Calcium" compartment="intravesicular" initialConcentration="0.36">
     <annotation>
       <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#" xmlns:bqbiol=http://biomodels.net/biology-qualifiers/
xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
         <rdf:Description rdf:about="#metaid_0000042">
           <bqbiol:isDescribedBy>
             <rdf:Bag>
               <rdf:li rdf:resource="urn:miriam:pubmed:12343565"/>
             </rdf:Bag>
           </bqbiol:isDescribedBy>
         </rdf:Description>
       </rdf:RDF>
     </annotation>
  </species>
```

Figure 4: Example of Core annotation applied to a species.

Using the current SBML annotation approach, it is not possible to annotate an attribute of an SBML element, such as the initial concentration of a species. The PubMed annotation in the example states that the `species` element as a whole is described by a particular PubMed reference (PubMed ID 12343565), while it was the intention to annotate the `species` attribute `initialConcentration`, effectively stating that the justification for the given initial concentration is described by the PubMed document with ID 12343565.

Statements about statements

With the current scheme all annotations of an SBML element are at the same level. They all relate to the element itself, but cannot be related to another statement. The ability to provide "statements about statements" is missing from Core annotation.

A simple use case is the request to annotate an annotation with the information that "this statement was added by...". A further use case would be annotations that involve non-binary relationships, such as "protein X is modified by modifier Y in position Z".

Relations between statements

In the Core annotation, it is currently not possible to define the relation between different annotations of a particular element. Apart from some conventions mentioned in the Core specification (see SBML L3 V1 Core Specification, p. 86) there is no fine-granular way of providing information on the annotation relations in a formal and specified manner.

The Core annotation standard syntactically limits the annotation of model constituents to:

```
<rdf:RDF>
 <rdf:Description rdf:about="#SBML_META_ID">
  <RELATION_ELEMENT>
   <rdf:Bag>
    <rdf:li resource="URI_1"/>
    <rdf:li resource="URI_2"/>
   </rdf:Bag>
  </RELATION_ELEMENT>
</rdf:RDF>
```

Figure 5: Current Core annotation syntax, illustrating the dependency on `rdf:Bag`.

RDF provides four different concepts to encode grouped statements, including the three Containers `rdf:Bag`, `rdf:Seq` and `rdf:Alt`, and the Collection `rdf:List` (see RDF Primer, sections 4.1 and 4.2):

- `rdf:Bag` represents an open group of resources or literals […] where there is no significance in the order of the members.
- `rdf:Seq` represents an open group of resources or literals [...] where the order of the members is significant.

- `rdf:Alt` represents an open group of resources or literals that are alternatives (typically for a single value of a property).
- `rdf:List` represents a closed group of resources or literates that consists only of the specified members.

The current Core annotation is restrictive as it does not allow the use of other containers than `rdf:Bag`, which only groups a set of statements, without implying any further semantics on the meaning of that group. Therefore, considering the above example, there is no way of currently determining what, if anything, the relationship is between `URI_1` and `URI_2`.

Examples of the ambiguity that can be caused by this limitation are highlighted in the following two examples. In the first example, the container `rdf:Bag` is used to define the relationship between two alternative annotations for glucose. In the second example, `rdf:Bag` is used to define the relationship between two components of a complex.

The first example effectively demonstrates an implied "or" relationship between two alternative means of annotating glucose (with a ChEBI term or a KEGG Compound term):

```
<species id=”glc" metaid=”meta_glc” name=“Glucose”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about=”#meta_glc”>
    <bqbiol:is>
     <rdf:Bag>
      <rdf:li rdf:resource="urn:miriam:obo.chebi:CHEBI%3417234"/>
      <rdf:li rdf:resource="urn:miriam:kegg.compound:C00234"/>
     </rdf:Bag>
    </bqbiol:is>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 6: Current Core annotation of species representing the simple molecule glucose.

`glc` is *either* `urn:miriam:obo.chebi:CHEBI:17234` **or** `urn:miriam:kegg.compound:C0023`, but not both.

The second example demonstrates an implied "and" relationship between two components of a complex (represented by a UniProt term for the protein, and a ChEBI term for the ligand):

```
<species id="Ca_calmodulin" metaid="cacam”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about="#cacam”>
    <bqbiol:hasPart>
     <rdf:Bag>
      <rdf:li rdf:resource="urn:miriam:uniprot:P62158"/>
      <rdf:li rdf:resource="urn:miriam:kegg.compound:C00076"/>
     </rdf:Bag>
    </bqbiol:hasPart>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 7: Current Core annotation of species representing the complex calcium calmodulin.

`Ca_calmodulin` has parts `urn:miriam:uniprot:P62158` **and** `urn:miriam:kegg.compound:C00076`.

The problem is that the relationship is implied: it is not made explicit by the container (`rdf:Bag`) used to define the relationship.

Furthermore, no clear definition of the different or similar meanings between using a list of

references in one `rdf:Bag` element as opposed to using a single `rdf:Bag` element for each reference is given. Consider the following two examples:

```
<rdf:RDF …>
 <rdf:Description rdf:about="#metaid_0000001">
  <bqbiol:is>
   <rdf:Bag>
    <rdf:li resource="x"/>
    <rdf:li resource="y"/>
   </rdf:Bag>
  </bqbiol:is>
 </rdf:Description>
</rdf:RDF>
```

```
<rdf:RDF …>
 <rdf:Description rdf:about="#metaid_0000001">
  <bqbiol:is>
   <rdf:Bag>
    <rdf:li resource="x"/>
   </rdf:Bag>
  </bqbiol:is>
  <bqbiol:is>
   <rdf:Bag>
    <rdf:li resource="y"/>
   </rdf:Bag>
  </bqbiol:is>
 </rdf:Description>
</rdf:RDF>
```

Figure 8: Current Core annotation examples indicating ambiguity between uses of different syntax to represent annotation with multiple resources.

Negative statements

The current Core annotation scheme does not allow for the definition of negative statements. That is, to make statements along the lines of "protein X is NOT phosphorylated".

Predicates and qualifiers

To satisfy RDF, predicates should be nouns, representing properties of the subject, rather than verbs as they are in the Core annotation. RDF triples should follow the pattern, "SUBJECT has PREDICATE whose value is OBJECT". Core annotations result in nonsensical RDF triples such as "SPECIES has IS_DESCRIBED_BY whose value is PUBMED:12345". It is proposed that the existing Biomodels.net predicates be updated, such that, taking the example above, "IS_DESCRIBED_BY" is replaced by "DESCRIPTION".

Doing so would allow the set of predicates (properties), and relationships between them, to be defined formally in an RDF schema (see http://www.w3.org/TR/rdf-primer/#rdfschema).

The Annot Package proposal

The following section summarises the proposals to be incorporated in the Annot package.

The examples enclosed within will use the proposed new predicates / qualifiers, as specified in the Appendix of this document.

Namespace and integration with SBML L3

The standard namespace for the Annot package is

`http://www.sbml.org/sbml/level3/version1/annot/version1`

A new version of the Annot package will be released with each new version of the Core package in order to comply with the new version of the Core (following the SBML L3 package

mechanism description).

In order to use the Annot package for SBML L3 models, the Annot namespace must be added to the `<sbml>` element namespace declarations:

```
<sbml xmlns="http://www.sbml.org/sbml/level3/version1/core" level="3"
version="1"
xmlns:annot="http://www.sbml.org/sbml/level3/version1/annot/version1"
...>
      ...
</sbml>
```

An SBML model can always be fully understood mathematically without understanding either the Core annotation or the Annot package extension annotation. Therefore, the use of the Annot package is optional. This can be defined by adding the XML attribute `annot:required` to the `sbml` element, and setting its value to `false`:

```
<sbml xmlns="http://www.sbml.org/sbml/level3/version1/core" level="3"
version="1"
xmlns:annot="http://www.sbml.org/sbml/level3/version1/annot/version1"
annot:required="false" ...>
      ...
</sbml>
```

Solutions

Statements about attributes

Sometimes, it is not only necessary to annotate an SBML element, but a more fine-grained annotation of a particular attribute of an element is needed.

The use of **XPath** (see http://www.w3schools.com/xpath/) to refer to a piece of XML inside the document is proposed. XPath is a standard technology for referencing elements and attributes inside an XML document, and it offers a well-defined scheme to do so. Furthermore, a great number of tools exist to evaluate XPath expressions.

Therefore, the `xpath` namespace is proposed, which allows the specification of any local object in the `rdf:about`:

```
rdf:about="xpath:XPathToTheObject"
```

One should use the element's id to refer to it, as in:

```
xpath://species[id='0001']/@initialConcentration
```

The following example shows an attribute annotation using the XPath notation.

```
 <species metaid="metaid_0000042" id="Y" name="Intravesicular
Calcium" compartment="intravesicular" initialConcentration="0.36">
    <annotation>
      <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#" xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
        <rdf:Description
rdf:about="xpath://species[id='Y']/@initialConcentration">
          <bqbiol:description>
            <rdf:Bag>
              <rdf:li rdf:resource="urn:miriam:pubmed:12343565"/>
            </rdf:Bag>
          </bqbiol:description>
        </rdf:Description>
      </rdf:RDF>
    </annotation>
 </species>
```

Figure 9: Example of proposed Annot annotation for annotation of attributes.

The recommended way of providing the XPath statement is to:

- Avoid addressing attributes and elements by their (ordering) number.
- Use the abbreviated syntax to identify an XML element in the model by its `id`, and then refer to the particular attribute.

It would be error-prone to use the XPath concept of addressing attributes and elements by their indices, as SBML does not support elemental ordering. As such, expressions along the lines of the example below are **not** recommended for use in the Annot package. Instead, the XPath statement should be specified by a reference to its `id`, as in the example given above.

✗ `//species[7]/@initialConcentration`

✓ `//species[id='Y']/@initialConcentration`

Secondly, whenever possible, instead of providing the full paths to elements or attributes, the abbreviated syntax should be used, which first selects all elements of the given element name from the SBML model, and then limits the result set depending on the given `id`. In XPath, a double forward slash (`//`) selects from all descendants of the context node as well as the context node itself. At the beginning of an XPath expression, it selects from all descendants of the root node. For example, the following XPath expression selects all `species` elements in the document:

```
//species
```

It is suggested that this syntax be used in the Annot package, given its simplicity in comparison to the more verbose syntax, which would entire the full path to be specified:

```
/sbml/model/listOfSpecies/species
```

Statements about statements

RDF Reification

RDF Reification, the standard method of making statements about statements, as described in the RDF Primer, section 4.3, will be utilised. This approach allows statements to be assigned to other statements that have an `rdf:ID` assigned. Subsequent statements refer to this statement by specifying the `rdf:ID` in the `rdf:about` attribute of the `rdf:Description` attribute.

The following example demonstrates Reification being used to make a statement about a statement:

```
<species id="abc" metaid="meta_abc">
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about="#meta_abc">
    <bqbiol:description rdf:ID="statement1">
     <rdf:Bag>
      <rdf:li rdf:resource="urn:miriam:pubmed:15387819"/>
     </rdf:Bag>
    </bqbiol:description>
   </rdf:Description>
   <rdf:Description rdf:about="#statement1">
    <dc:creator>John Smith</dc:creator>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 10: Example of RDF Reification.

By adding an `rdf:ID` to the first statement (which states that the `species` has description PubMed document 15387819), a second statement can be specified about this first statement, which specifies that the first statement has a specified creator. Effectively the

second statement defines that the first statement has creator John Smith.

Person's meta-annotations

Core annotations are limited regarding specifying information on the different people involved in the model building, publishing, curating and maintaining process. In the Annot package, the use the `dc:creator` from Dublin Core to provide meta-information about persons is again proposed. However, it is proposed that such an annotation can be applied to both the `model` itself and any of the `model` sub-elements.

It is assumed that such annotations are inherited from parent nodes when a given node is not annotated with a `dc:creator`. For example, if a `model` element is annotated with a `dc:creator` but none of its sub-elements are, it is assumed that all sub-elements have been created by the `model` creator.

Non-binary relations

Related to this is the support for capturing non-binary relationships through the utilisation of blank nodes.

This example captures the statement "Hexokinase 2 is modified by phosphoserine in position 158", by specifying a blank node (`node1`) as the **object** of the modification **predicate**, and utilising this blank node as the **subject** of two subsequent statements. Note that phosphoserine is represented by the MIRIAM URN `urn:miriam:obo.psi-mod:MOD%3A00046`.

```
<species id=”x" metaid=”meta_x” name=“Hexokinase 2”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about=”#meta_x”>
    <bqbiol:modification rdf:nodeID=”node1"/>
   </rdf:Description>
   <rdf:Description rdf:nodeID=”node1”>
    <bqbiol:modifier rdf:resource=”urn:miriam:obo.psi-
mod:MOD%3A00046"/>
    <bqbiol:position rdf:datatype="xsd:integer">158</bqbiol:position>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 11: Example of the use of blank nodes.

Essentially, this specifies the following three triples:

meta_x	bqbiol:modification	node1
node1	bqbiol:modifier	urn:miriam:obo.psi-mod:MOD%3A00046
node1	bqbiol:position	158

Relations between statements

To enable a more detailed description of relations between statements, it is proposed to extend the current SBML annotation scheme to support all RDF Collections and Containers (`rdf:Bag`, `rdf:Seq`, `rdf:Alt`, and `rdf:List`).

The Core annotations specify that an `rdf:Bag` must be used. This, however, is unnecessary for single objects that can be specified more simply following the example syntax below:

```
<species id=”glc" metaid=”meta_glc” name=“Glucose”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about=”#meta_glc”>
    <bqbiol:identity
rdf:resource="urn:miriam:obo.chebi:CHEBI%3417234"/>
   </rdf:Description>
```

```
  </rdf:RDF>
 </annotation>
</species>
```

Figure 12: Example of valid RDF/XML that does not use either a Collection or Container.

In addition to supporting all RDF Collections and Containers, the use of no Collections and Containers will be supported. Considering all RDF Collections and Containers, and taking the previous examples (see Problems with Core annotation: Relations between statements), the existing, Core annotation implied "or" relationship between two alternative means of annotating glucose (with a ChEBI term or a KEGG Compound term) can be made explicit by using the `rdf:Alt` collection:

```
<species id=”glc" metaid=”meta_glc” name=“Glucose”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about=”#meta_glc”>
    <bqbiol:identity>
     <rdf:Alt>
      <rdf:li rdf:resource="urn:miriam:obo.chebi:CHEBI%3417234"/>
      <rdf:li rdf:resource="urn:miriam:kegg.compound:C00234"/>
     </rdf:Alt>
    </bqbiol:identity>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 13: Example of using the Container `rdf:Alt` to represent an "or" relationship between resources.

Similarly, the existing, Core annotation implied "and" relationship between two components of a complex (represented by a UniProt term for the protein, and a ChEBI term for the ligand) can be made explicit by utilising the `rdf:List` collection to specify a closed set:

```
<species id="Ca_calmodulin" metaid="cacam”>
 <annotation>
  <rdf:RDF>
   <rdf:Description rdf:about="#cacam”>
    <bqbiol:part>
     <rdf:List>
      <rdf:li rdf:resource="urn:miriam:uniprot:P62158"/>
      <rdf:li rdf:resource="urn:miriam:kegg.compound:C00076"/>
     </rdf:List>
    </bqbiol:part>
   </rdf:Description>
  </rdf:RDF>
 </annotation>
</species>
```

Figure 14: Example of using the Collection `rdf:List` to represent an "and" relationship between resources.

Distinction between L3 Core and L3 Annot package annotations

To distinguish SBML Level 3 Core annotations from annotations provided through the Annot package, a new element `<annot:annotation>` from the `annot` namespace as a sibling of the current `<annotation>` element is proposed. This will allow L3 Annot package annotations, i.e. the ones in the scope of this draft proposal, and existing Core annotations to be distinguished.

The following example shows the `annot:annotation` element as a sibling of the current SBML annotation element:

```
<annotation>
      [CORE ANNOTATION]
```

```
</annotation>
<annot:annotation>
      [ANNOT PACKAGE ANNOTATION]
</annot:annotation>
```

The approach chosen here has the advantage of this approach is that it avoids further overloading of the already much used Core `annotation` element. It also allows a cleaner distinction between the Core and Annot package annotations.

The recommended practice for model annotation is the use of the Annot package, as it is less restricted in its syntax, and complies with RDF recommendations.

Cross-references and cross-element annotations

Self-references

In order to realise self-references, i.e. to refer to an element in the same document, use of the existing RDF standard will be supported:

```
<rdf:li rdf:resource="#metaID">
```

Non-URI references

The referencing of non-URI references to existing models (such as the example below), such as web addresses, URLs, or local directories, is NOT supported by this proposal.

```
<rdf:li rdf:resource="file://../models/BM02#_986127"/>
```

Negative statements

No suitable solution has been proposed for specifying negative statements. This issue will be addressed in a subsequent iteration of the Annot package.

Predicates and qualifiers

It is recognised that the current set of predicates (Biomodels.net qualifiers, http://www.ebi.ac.uk/miriam/main/qualifiers/) should be extended the RDF specification that predicates should be nouns, representing properties of the subject, rather than verbs as they are in the Core annotation (see http://www.w3.org/TR/rdf-primer/#rdfschema).

This process will be delegated to the developers of Biomodels.net. An initial mapping of existing (verb) predicates to new (noun) predicates is available in the Appendix.

Package dependencies

This package does not depend on any other SBML Level 3 package.

Prototype implementations

No prototype implementation exists as yet.

Translation to SBML Level 2

Translation of Annot package annotations back to SBML Level 2 annotations will not be supported.

Hints

Use of the Annot package

There is no way to legislate how other packages make use of the Annotation structures coming from this package. Individual packages determine how best to make use of Annotation structures.

Use of old and new annotations

Duplicating semantic information (in both Core annotation and the Annot package annotation) is technically possible, but it is considered bad practice and not recommended. Instead, it is recommended that, if Annot package annotation is to be used, these annotations should replace any existing Core annotations within the model.

Appendix

Predicates and qualifiers

It is recognised that the current set of predicates (Biomodels.net qualifiers, http://www.ebi.ac.uk/miriam/main/qualifiers/) should be extended the RDF specification that predicates should be nouns, representing properties of the subject, rather than verbs as they are in the Core annotation (see http://www.w3.org/TR/rdf-primer/#rdfschema).

This process will be delegated to the developers of Biomodels.net.

It is intended that the new predicates will coexist with the existing predicates, with the recommendation that the new set be used in preference to the existing set.

The following describes the mapping between old and new predicates. Where multiple options exist, these indicate candidate predicates that will be decided by the Biomodels.net community.

`bqmodel:is`	`bqmodel:identity`
`bqmodel:isDerivedFrom`	`bqmodel:progenitor, bqmodel:antecedent, bqmodel:ancestor, bqmodel:basis, bqmodel:base, bqmodel:foundation, bqmodel:origin`
`bqmodel:isDescribedBy`	`bqmodel:description`
`bqbiol:hasPart`	`bqbiol:part`
`bqbiol:hasProperty`	`bqbiol:property`
`bqbiol:hasVersion`	`bqbiol:version`
`bqbiol:is`	`bqbiol:identity`
`bqbiol:isDescribedBy`	`bqbiol:description`
`bqbiol:isHomologTo`	`bqbiol:homolog`
`bqbiol:isEncodedBy`	`bqbiol:encoder`
`bqbiol:encodes`	`bqbiol:encodement`
`bqbiol:isPartOf`	`bqbiol:encompassment, bqbiol:assembly, bqbiol:partship, bqbiol:parthood, bqbiol:whole, bqbiol:meronym`
`bqbiol:isPropertyOf`	`bqbiol:bearer, bqbiol:carrier`
`bqbiol:isVersionOf`	`bqbiol:consociate, bqbiol:cohort, bqbiol:superclass, bqbiol:hyponym`
`bqbiol:occursIn`	Physical containment: `bqbiol:encompassment, bqbiol:containment`
`bqbiol:occursIn`	Taxonomic instantiation: `bqbiol:instantiation`

Publication 6

Enzyme kinetics informatics: from instrument to browser.

Swainston N, Golebiewski M, Messiha HL, Malys N, Kania R, Kengne S, Krebs O, Mir S, Sauer-Danzwith H, Smallbone K, Weidemann A, Wittig U, Kell DB, Mendes P, Müller W, Paton NW, Rojas I.

FEBS J. 2010, **277**, 3769–79.

the FEBS Journal

Enzyme kinetics informatics: from instrument to browser

Neil Swainston[1,*], Martin Golebiewski[2,*], Hanan L. Messiha[1], Naglis Malys[1], Renate Kania[2], Sylvestre Kengne[2], Olga Krebs[2], Saqib Mir[2], Heidrun Sauer-Danzwith[2], Kieran Smallbone[1], Andreas Weidemann[2], Ulrike Wittig[2], Douglas B. Kell[1], Pedro Mendes[1,3], Wolfgang Müller[2], Norman W. Paton[1] and Isabel Rojas[2]

1 Manchester Centre for Integrative Systems Biology, University of Manchester, UK
2 Heidelberg Institute for Theoretical Studies, Germany
3 Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

Keywords
data analysis; database; enzyme; kinetics; metadata

Correspondence
N. Swainston, Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester M1 7DN, UK
Fax: +44 161 306 8918
Tel: +44 161 306 5146
E-mail: neil.swainston@manchester.ac.uk
Website: http://www.mcisb.org

*These authors contributed equally to this work

(Received 31 May 2010, revised 20 June 2010, accepted 13 July 2010)

doi:10.1111/j.1742-4658.2010.07778.x

A limited number of publicly available resources provide access to enzyme kinetic parameters. These have been compiled through manual data mining of published papers, not from the original, raw experimental data from which the parameters were calculated. This is largely due to the lack of software or standards to support the capture, analysis, storage and dissemination of such experimental data. Introduced here is an integrative system to manage experimental enzyme kinetics data from instrument to browser. The approach is based on two interrelated databases: the existing SABIO-RK database, containing kinetic data and corresponding metadata, and the newly introduced experimental raw data repository, MeMo-RK. Both systems are publicly available by web browser and web service interfaces and are configurable to ensure privacy of unpublished data. Users of this system are provided with the ability to view both kinetic parameters and the experimental raw data from which they are calculated, providing increased confidence in the data. A data analysis and submission tool, the KINETICS-WIZARD, has been developed to allow the experimentalist to perform data collection, analysis and submission to both data resources. The system is designed to be extensible, allowing integration with other manufacturer instruments covering a range of analytical techniques.

Introduction

The field of systems biology is heavily reliant on reliable experimental data in order to create predictive models. With the establishment of high-throughput technologies in genomics, proteomics and metabolomics over the past decade, the amount of data available to the biochemistry community is increasing exponentially. However, the collection and dissemination of experimental data can be a labour-intensive process, such that much acquired data never becomes available to the community in an accessible and utilizable form. Thus, the data flow from the experiment to the consumer performing the analysis, the comparison or the set-up of computer models can still constitute a bottleneck. This problem calls for systems that capture the data directly from the experimental instrument, process and normalize it to agreed standards and finally transfer these data to publicly available databases to make them accessible.

To facilitate the dissemination of data, a number of initiatives have been developed to advise on the minimum requirements to follow in the storage and dissemination of experimental data in fields such as transcriptomics and proteomics, which will ultimately allow data to be easily and freely shared between

Abbreviations
SBML, Systems Biology Markup Language; SBRML, Systems Biology Results Markup Language; STRENDA, Standards for Reporting Enzymology Data; XML, Extensible Markup Language.

laboratories worldwide [1–3]. The enzyme kinetics community is active in this area with the development of both standardized experimental operating procedures [4] and recommendations on data storage in the form of the Standards for Reporting Enzymology Data (STRENDA, http://www.strenda.org) [5] guidelines.

There exist several publicly available databases containing enzyme kinetics data that adhere to these recommendations, with BRENDA [6], a database for enzyme functional data, and the biochemical reaction kinetics database, SABIO-RK [7], as the two most comprehensive and most used examples. The data gathered in these resources were typically manually extracted from the biochemical literature and entered into the databases by hand, a labour-intensive and time-consuming process. To support this manual work SABIO-RK offers a tailored input interface [8], which allows users to manually enter kinetic data and corresponding metadata, utilizing standardized terms in the form of controlled vocabularies and references to external resources. BRENDA recently also introduced the support of kinetic parameter submission. These input interfaces could in principle assist experimenters in submitting their kinetic data to the databases. However, entering each dataset manually can be a tedious and error-prone process and is unlikely to be accepted as standard practice by the scientific community. To date there has been no support for automated submission of kinetic data or for storage of the original raw experimental data from which these constants were calculated.

We here introduce an automated system to support the whole workflow of deriving kinetic data from the laboratory instrument and make it accessible in the web. The task of managing enzyme kinetics data involves four steps: data capture, analysis, submission and querying/visualization. The first three tasks have been integrated in a unified tool, the KINETICSWIZARD. Data querying and visualization are provided by web browser interfaces for manual access and web services for automated access to both the newly developed MeMo-RK and the existing SABIO-RK databases.

The KINETICSWIZARD, introduced here, provides a unified interface for capturing and fitting raw kinetics time series data along with sufficient metadata to allow these data to be queried, such as detailed and unambiguous descriptions of the reactions studied, their reactants and modifiers, and experimental conditions. Data can then be automatically submitted to the relevant data repositories. By collecting this in a principled manner, the intention is that any data collected and submitted to the repositories will be complete, consistent and adhere to defined standards, such as the STRENDA recommendations.

Although much of this work has been developed in the context of systems biology, the tools described are sufficiently generic to be used in other fields, such as molecular enzymology and drug discovery.

Results

Data capture, analysis and submission

KINETICSWIZARD data capture

The key to ensuring that resources storing enzyme kinetics data can be usefully employed in a systems biology environment is in the richness and accuracy of the metadata associated with the kinetic constants. Specifically, for instance, we need to know the experimental conditions under which *in vitro* assays were performed, such as pH, temperature and buffer. Additionally, the components of the assay, such as enzyme variants, substrates, products and modifying molecules, must be unambiguously defined.

Designed to be used by experimentalists rather than bioinformaticians, the KINETICSWIZARD is intended to hide much of the more technical aspects of data management from the user and present an intuitive, user-friendly interface from which this necessary metadata can be obtained. The KINETICSWIZARD can be launched automatically from the instrument software, allowing data to be captured, analysed and submitted to databases immediately upon acquisition. The KINETICSWIZARD has been developed initially to integrate with a BMG Labtech NOVOstar instrument (Offenburg, Germany). A generic version, which reads experimental data from a spreadsheet, along with an example of experimental data in this spreadsheet format, is also available (http://www.mcisb.org/resources/kinetics/). The system has been designed in a modular manner to allow the support of different instruments and experimental techniques (see Fig. 1).

In a typical experimental set-up, the user runs several time series assays, in which a reactant concentration is varied. The WIZARD allows the user to specify these varying reactant concentrations, which are then associated with the experimental data, and used in the subsequent fitting step to calculate kinetic parameters. A number of assays can be 'grouped' together, supporting experimental set-ups in which numerous reactions are assayed on a single plate.

To provide this functionality, the tool draws heavily on the use of existing data resources that are relevant to the task, and queries these resources via web service interfaces where possible. Exploiting existing data resources has the advantages of greatly reducing the volume of metadata that the experimentalist must

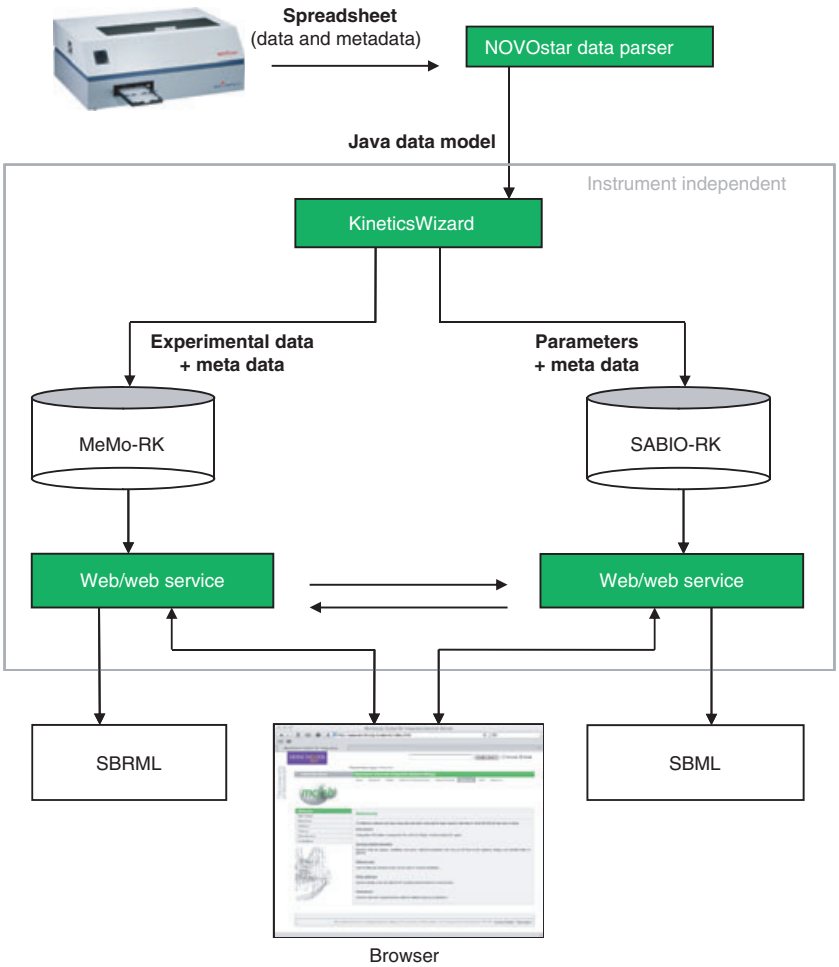

Fig. 1. Enzyme kinetics from instrument to browser. Data are extracted from the NOVOstar instrument as a Microsoft Excel spreadsheet. They are parsed into a data model and imported into the KINETICSWIZARD. The KINETICSWIZARD provides a graphical user interface that allows the experimenter to associate metadata to the experimental data. Kinetic constants are then calculated and the data submitted to appropriate repositories: MeMo-RK (http://www.mcisb.org/MeMo-RK/) for the experimental raw data, and SABIO-RK (http://sabio.h-its.org/) for the derived kinetic parameters, equations and appropriate metadata. Links are maintained between the repositories allowing both raw data and parameter sets to be accessed through web browser interfaces and web services. Kinetic data can be exported from SABIO-RK in SBML format and experimental data exported from MeMo-RK in SBRML format.

submit, while also annotating the submitted data with standard ontological terms to facilitate subsequent querying.

An example is in the specification of the reaction itself. It has been reported that relying on textual descriptions of small molecules and enzymes can result in inconsistencies, as the naming of such species is largely subjective and can differ greatly from individual to individual [9]. The KINETICSWIZARD ensures the consistent specification of reaction components by utilizing libAnnotationSBML [10], a library that provides an interface to the KEGG database [11]. The user specifies an organism and a gene name, from which the KEGG web service is queried and all reactions catalysed by the enzyme encoded by the supplied gene are returned (see Fig. 2). An individual reaction can then be selected, and the corresponding database entry queried to harvest a number of terms that would otherwise have to be specified manually by the user, such as EC term and the identity of reactants, products and enzyme. By utilizing KEGG reactions in this way, reaction participants are specified internally as entries in either the KEGG or the ChEBI [12] databases, and enzymes as UniProt [13] terms. Accurate stoichiometry of each of the reaction participants is also gathered. This provides an unambiguous, computer-readable 'signature' for the specified reaction, which facilitates subsequent querying of the data themselves.

Situations may arise in which reactions are being studied that are not in the KEGG database. Future iterations could query other sources containing such data, such as Reactome [14], BRENDA or SABIO-RK itself. Alternatively, the user interface could be extended to allow the user to specify the reaction manually. However, this approach would put a greater burden on the user, and would increase the likelihood that inconsistent reactants, enzymes, EC terms, etc., would be input.

After defining the reaction, the user is provided with the facility to specify buffer reagents and coupling enzymes, along with other metadata values, including the environmental conditions, such as pH and temperature, under which the assays were performed.

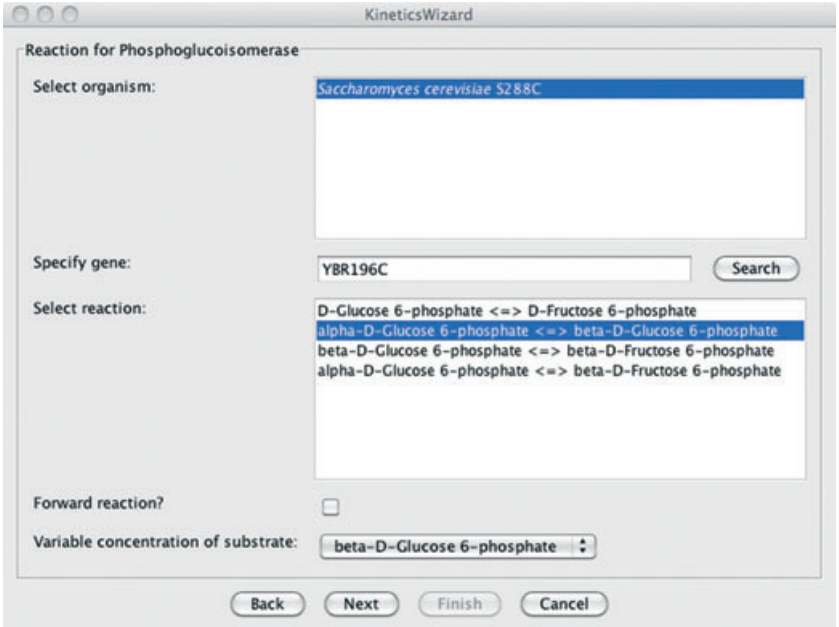

Fig. 2. Specifying the reaction components. Upon specification of an organism and a gene, a search is performed against the KEGG web service, allowing the user to select from a list of reactions. The user can then specify the direction of the reaction, and which substrate concentration was varied during the assays.

In order to ensure that a given parameter is used in the intended manner, it is also necessary to specify the kinetic mechanism and equation that was used to determine the parameter. The initial version of the KINETICSWIZARD assumes that all reaction mechanisms follow irreversible, steady-state Henri–Michaelis–Menten kinetics [15]. Future releases of the KINETICSWIZARD will support more complex mechanisms, for example in cases where inhibition or allostery is observed. The kinetic mechanism and all kinetic parameters are specified internally, and later archived with, unambiguous terms from the Systems Biology Ontology [16].

Utilizing existing bioinformatics resources provides the twin advantages of reducing the burden on the experimentalist in redefining metadata that are already present digitally elsewhere, while also ensuring the consistency of the metadata, aiding subsequent comparisons, analyses and reuse of data from different experiments or different laboratories.

v_{max} parameters are often specified without any indication of the enzyme concentration contained within the term. To prevent this, the KINETICSWIZARD captures the enzyme concentration used in the assay, allowing the kinetic parameter to be submitted as a k_{cat} value. This decouples the parameter from the enzyme concentration and increases the usability of the value. To facilitate this further, standard units are specified for all parameters, with substrate and enzyme concentrations input in mM and nM, respectively.

Finally, a free text field is available, allowing the user to assign notes and comments to the dataset.

KINETICSWIZARD data analysis

Following the data capture phase, the next step before data submission is data analysis [17], whereby kinetic parameters are determined by applying an appropriate fitting algorithm to the experimental time series data. By default, the initial version of the KINETICSWIZARD provides a fitting algorithm that assumes irreversible Henri–Michaelis–Menten kinetics. As the tool develops further, fitting to other more complex kinetic mechanisms will be supported.

During the experimental set-up, individual assays may be specified as being either samples or blanks. Blanks are assays that contain all components apart from the enzyme under investigation, and if present their data are subtracted from those of the sample assays. A straight-line fit is then used to estimate initial reaction rates. These values are then fed into the Eadie–Hofstee linearized version of the Michaelis–Menten equation [18,19] to provide estimates of k_{cat} and K_M. More accurate parameter values are subsequently obtained through nonlinear regression via the Levenberg–Marquardt algorithm [20,21]. Although the curve-fitting algorithm is automated, the user is provided with a visual representation of the fit from which the initial rate is calculated. The user may then perform a manual refit by dragging the initial rate line; a feature that can be utilized to correct for lag times of coupling enzymes, for example. Overriding the automated initial rate calculation will update the calculated k_{cat} and K_M parameters in real time (see Fig. 3).

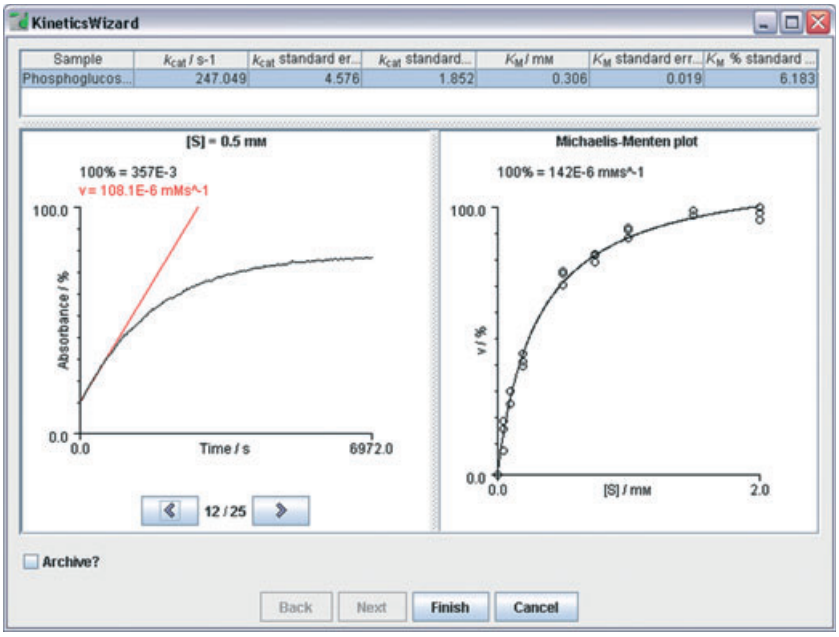

Fig. 3. Displaying and manipulating the results of the curve-fitting algorithm. The left-hand panel allows the user to view each assay in the data set and its automatically fitted initial rate. The red initial-rate line may be manually corrected by dragging, allowing the default fit to be overridden for noisy or anomalous data. These initial rates are plotted against substrate concentration in the right-hand panel, which shows the Michaelis–Menten curve. The top panel shows the calculated kinetic parameters k_{cat} and K_M, together with their standard errors. Manually correcting an initial rate updates both the Michaelis–Menten curve and the calculated kinetic parameters in real time.

In order to test and validate the KINETICSWIZARD fitting algorithm, home-produced enzymes have been assayed (see Materials and methods). A number of time series assays were acquired for each enzyme, and the data captured and analysed using the KINETICSWIZARD. The calculated kinetic parameters were comparable with those calculated by the GRAFIT software package (Erithacus Software Ltd, Horley, UK), version 5.0 (see Table 1).

KINETICSWIZARD submission tool

The data submission task is two-fold: submission of the raw experimental data to MeMo-RK and submission of derived kinetic equations with their kinetic parameters and corresponding metadata to SABIO-RK.

MeMo-RK is a derivation of the MeMo database, originally constructed for storage of metabolomics data [22]. It has been amended to store raw, experimental kinetics data and associated metadata, including submitter, laboratory, instrument settings and experiment type, such as absorbance or fluorescence.

Derived, secondary data in the form of kinetic parameters and equations, definitions of the reactions being studied and relevant metadata describing the experimental and environmental conditions such as temperature, pH, buffer solution, coupling enzymes are represented in an Extensible Markup Language (XML) document and submitted directly to the SABIO-RK submission web service. SabioML, a novel XML schema, has been developed for this purpose and could also serve as a kinetic data transfer format between sources other than SABIO-RK. Derived from the SABIO-RK database schema [23], it comprises kinetic laws, parameters and relevant metadata in a structured and standardized format, exploiting a controlled vocabulary and appropriate

Table 1. Comparison of kinetic parameters calculated by the KINETICSWIZARD and GRAFIT. Detailed views of the reaction, parameters and metadata can be found at the appropriate SABIO-RK records, http://sabio.h-its.org/kineticLawEntry.jsp?kinlawid=29371, http://sabio.h-its.org/kineticLawEntry.jsp?kinlawid=29401 and http://sabio.h-its.org/kineticLawEntry.jsp?kinlawid=29390, respectively).

Enzyme	KINETICSWIZARD	GRAFIT
Fructose-bisphosphate aldolase (ALF1_YEAST, EC: 4.1.2.13)	k_{cat}: 4.14 ± 0.061 s^{-1} K_M: 0.451 ± 0.024 mM	k_{cat}: 4.27 ± 0.097 s^{-1} K_M: 0.442 ± 0.037 mM
Pyruvate decarboxylase isozyme 2 (PDC5_YEAST, EC: 4.1.1.1)	k_{cat}: 1.78 ± 0.037 s^{-1} K_M: 11.4 ± 0.65 mM	k_{cat}: 1.79 ± 0.029 s^{-1} K_M: 11.3 ± 0.51 mM
Glucose-6-phosphate isomerase (G6PI_YEAST, EC: 5.3.1.9)	k_{cat}: 247 ± 5.1 s^{-1} K_M: 0.307 ± 0.021 mM	k_{cat}: 253 ± 5.1 s^{-1} K_M: 0.304 ± 0.020 mM

ontologies. Upon submission, the data are held in a gatekeeper database that can only be accessed by the submitter and curators of SABIO-RK. Upon formal curation and release by the submitter, the data are then made public in the database. This process ensures consistency and completeness of the data and provides data confidentiality, such that data can remain privately accessible before publication.

The KINETICSWIZARD can be configured to perform these submission steps automatically, ensuring that both experimental data and derived kinetic parameters are captured and stored immediately upon acquisition and analysis.

Data access

Access to the submitted data utilizes the two data repositories, MeMo-RK for experimental raw data and SABIO-RK for derived kinetic equations with their parameters and corresponding metadata. This approach is consistent with a distributed, loosely coupled system [24], in which a number of independent data resources are populated, and then later queried via web browser or web service interfaces. The key to the development of such a distributed system is to ensure a consistent means of identifying species, reactions and parameters across each of these data resources. Data submitted from the KINETICSWIZARD populates both databases, and from this, each resource can subsequently cross-reference the other, providing a link from kinetic parameters to raw data and vice versa.

An advantage of this approach is that it uncouples the storage of raw data from the storage of derived kinetic parameters, such that users have a single interface to query and retrieve kinetic parameters, irrespective of whether they have been extracted from literature or submitted by the KINETICSWIZARD. Also, this separation facilitates submission of kinetic parameters to other repositories, such as BRENDA, without affecting the raw data storage in MeMo-RK.

Web browser interface

Both MeMo-RK and SABIO-RK have web browser interfaces. SABIO-RK provides an interface for performing sophisticated searches for kinetic parameters, based on a combination of reactants, enzymes, organisms, tissues, pathways, experimental conditions, etc. Pages displaying a set of kinetic parameters link to the original data source, e.g. to the PubMed reference of the paper from which the data have been extracted, or to the corresponding page in MeMo-RK displaying the raw experimental data where the data have been submitted from the KINETICSWIZARD (see Table 1 and Fig. 4). Similarly, MeMo-RK provides a link to the associated kinetic parameters in SABIO-RK, and contains a searchable interface to the raw experimental data (see Fig. 5).

Web service interface

The SABIO-RK web services (http://sabio.h-its.org/webservice.jsp) provide flexible programmatic access to the data, allowing users to write clients to customize and automate access directly from their simulation software, systems biology platforms, tools or databases [25]. The web services provide customizable points of entry and thereby an extensive search capability for kinetic data and corresponding metadata stored in SABIO-RK. The task of automatically finding parameters and associated data is aided by specifying and storing metadata using controlled vocabularies and ontological terms. As in the web browser interface, reactions with their kinetic data can be exported in Systems Biology Markup Language (SBML) [26]. An example of direct access to kinetic data through these web services has been implemented in CELLDESIGNER, a modelling tool for biochemical networks [27].

Once a given set of kinetic parameters has been discovered from the SABIO-RK web services, the user may then retrieve associated raw data in Systems Biology Results Markup Language (SBRML) [28] format via the MeMo-RK web services, allowing the data to be viewed or refitted. Such a query across distributed web services can be performed with specialized workflow software, such as TAVERNA [29].

Discussion

The development of this system was driven by the need to exchange kinetic data between experimentalists and consumers, particularly in the context of high-throughput assays and the integration of their results into biochemical computer models for simulation. Such a system had the following requirements: to provide a means of calculating kinetic parameters from raw experimental data; to store these parameters in a standardized and consistent way, such that they can readily be queried and used in systems biology studies [30,31]; and to archive the raw experimental data such that it could be reused if required, e.g. for quality control or for refitting. Furthermore, the system was to be applicable to data from a number of instruments using different experimental techniques, and the intended users of the system were experimental biologists, not bioinformaticians.

General information	
Organism	Saccharomyces cerevisiae
Strain	BY4700 transformed in Y258
Tissue	-
EC Class	5.3.1.9
SABIO reaction id	1123
Variant	wildtype Glucose-6-phosphate isomerase
Recombinant	expressed in Saccharomyces cerevisiae

Substrates		
name	location	comment
D-Fructose 6-phosphate	-	-

Products		
name	location	comment
D-Glucose 6-phosphate	-	-

Modifiers				
name	location	effect	comment	protein complex
Glucose-6-phosphate isomerase(Enzyme)	-	Modifier-Catalyst	-	(P12709)*2;

Enzyme (protein data)				
	UniProt-ID	name	mol. weight (kDa)	deviation (kDa)
subunit	-	-	-	-
complex	-	-	-	-

Kinetic Law	
type	formula
Michaelis-Menten	k_{cat} * Et * S / (Ks + S)

Parameter							
name	type	species	start val.	end val.	deviat.	unit	comment
k_{cat}	k_{cat}	-	247.20936123	-	5.0661300497	s^(-1)	-
Et	concentration	Enzyme	6.666666E-7	-	-	mM	-
K_S	K_m	D-Fructose 6-phosphate	0.30698334365	-	0.020968017498	mM	-
S	concentration	D-Fructose 6-phosphate	-	2	-	mM	-

Experimental conditions			
	start value	end value	unit
temperature	30	-	°C
pH	6.5	-	-
buffer	0.0 mM Glucose-6-phosphate 1-dehydrogenase, 100.0 mM 2-(N-morpholino)ethanesulfonic acid, 5.0 mM magnesium dichloride, 100.0 mM potassium chloride, 0.4 mM NADP		

Reference				
SABIORK id	title	author	year	direct submission link
2459	Glycolysis	Hanan Messiha and Naglis Malys, Manchester Centre for Integrative Systems Biology (MCISB), UK	2009	http://beaconw.cs.manchester.ac.uk:8092/mcisb-web/index.jsp?application=MeMo-RK&file=index.jsp&directory=Browse&homepage=index.jsp&experimentId=_14795743_9a81_45eb_ab79_d779a61fa41b

Fig. 4. Screen capture of the web browser interface to SABIO-RK (http://sabio.h-its.org/), showing a coherent set of kinetic parameters submitted from the KINETICSWIZARD. A cross-link to the corresponding experimental raw data in MeMo-RK is shown at the bottom.

The KINETICSWIZARD addresses many of these issues by providing an interactive tool that integrates with instrumentation software and allows kinetic parameters to be calculated from experimental data, also providing the facility to manually correct the automated fit for noisy or anomalous data. The data model representing raw experimental data is a simple one that can be applied to many experimental techniques.

The tool manages the collection of metadata and the submission of these data to appropriate resources. In order to facilitate both the querying of these resources and subsequent data integration, standardized terms or references to external resources are associated with the data, and these can be assigned in an intuitive, user-friendly manner. Considering systems biology studies, the task of parameterizing models with kinetic parameters is greatly simplified with data in this form, as both the SBML file containing the model and the underlying data stored in the resources can be annotated with the same terms for metabolites, enzymes, EC codes, parameter types, etc. This task is facilitated by the storage of kinetic data in SABIO-RK, from which data

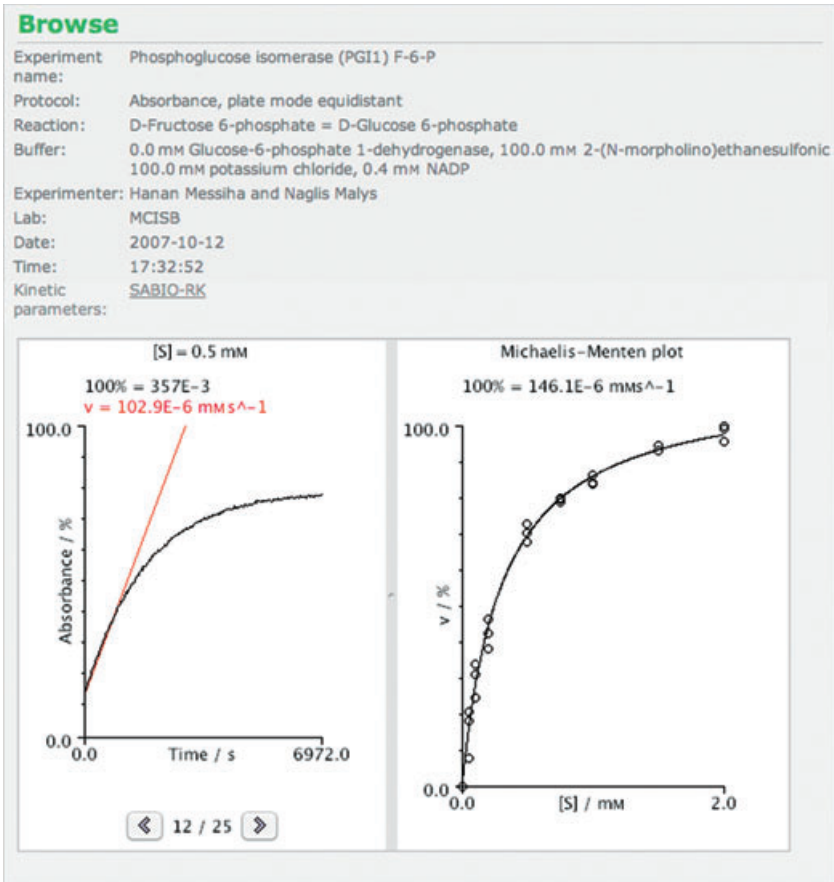

Fig. 5. Screen capture of the web browser interface to MeMo-RK (http://www.mcisb.org/MeMo-RK/), showing instrument raw data, the Michaelis–Menten curve and a link to parameter data in SABIO-RK.

can be exported in SBML format either through a web browser or web service interface.

Beyond the calculation, storage and dissemination of kinetic parameters, another primary focus of the work is on the management and distribution of raw experimental data. It is hoped that the introduction of a system for the storage and retrieval of raw enzyme kinetics assay data will encourage the community to share such data and to make it available in tandem with any kinetic parameters that are published. The proteomics community have made progress in this area in recent years, both with the development of standards for representing data [32] and encouraging major journals to advise that instrument data be shared in addition to derived results [33,34]. Crucially, such efforts have been supported by the development of software tools to aid experimentalists in making their data available [35–37]. It is hoped that the introduction of such a system here, along with the standardization efforts of the STRENDA commission, will encourage comparable behaviour in the enzyme kinetics community, such that the publication of enzyme kinetic parameters without the sharing of associated experimental data becomes the exception rather than the norm.

Materials and methods

Enzyme expression, purification and quantification

Enzymes were expressed in *Saccharomyces cerevisiae* strains containing either overexpression plasmid [38] or chromosome-integrated gene fusion [39] and purified essentially as described previously [40]. Enzyme purity was analysed by SDS/PAGE according to Laemmli [41]. The amount and concentration of purified enzyme was determined using a standard method [42] and preparation quality confirmed with the 2100 Bioanalyzer (Agilent Technologies, Foster City, CA, USA).

Kinetic assays

Kinetic time course data of purified enzymes were determined in high-throughput measurements using a NOVOstar plate reader in 384-well format plates. All measurements were carried out at 30 °C in 60 μL reaction volumes in a reaction buffer that consisted initially of 100 mM Mes, pH 6.5, 100 mM KCl and 5 mM free magnesium chloride plus other reagents and substrates that were specific for each individual enzyme.

Assays were automated so that all reagents in the reaction buffer were in 45 μL, enzyme in 5 μL and substrate in 10 μL volumes. In almost all cases, the enzyme was incubated in the reaction mixture and the reactions were started by the addition of the substrate.

Assays for each individual enzyme were either developed or modified from previously published methods to be compatible with the conditions of the reactions (e.g. pH compatibility or unavailability of commercial substrates). For each individual enzyme, the forward and the backward reaction were measured whenever applicable, depending on the possibility of the production of active enzyme, the availability of substrates as well as the suitability of the assays at the specified pH. Some assays were modified, altering the concentration of coupling enzymes or other reagents to ensure that the rate measured was the rate of the reaction of interest.

All assays were coupled to enzymes where NAD(P) or NAD(P)H was a product or substrate whose formation or consumption could be followed spectrophotometrically at 340 nm using an extinction coefficient ($\Sigma_{340\ nm}$) of 6.620 $mM^{-1} \cdot cm^{-1}$.

All measurements were based on at least duplicate determination of the reaction rates at each substrate concentration. For all assays, control experiments were run in parallel to correct for any unwanted background activity.

Implementation and distribution

The KINETICSWIZARD, MeMo-RK web browser interface and web service interface are written in JAVA 1.6. MeMo-RK has been tested on POSTGRESQL 8.3. All are supported in Windows and MacOS X and are distributed as source code and associated build files. They are distributed under the open source Academic Free Licence v3.0 from http://mcisb.sourceforge.net. An example version of the KINETICSWIZARD, and usage instructions, can be found at http://www.mcisb.org/resources/kinetics/, together with links to the MeMo-RK web browser and web service interfaces. The SABIO-RK web browser and web service interfaces, submission tool and the transfer procedures are written in JAVA 1.6 and owned by HITS gGmbH (Heidelberg Institute of Theoretical Studies, Heidelberg, Germany). The SABIO-RK database system is currently implemented in Oracle 10 g and is owned by HITS gGmbH. Free access to data in SABIO-RK is granted for academic use via web browser interface or web services. Terms and conditions can be found at the SABIO-RK homepage (http://sabio.h-its.org/).

Acknowledgements

The authors thank the EPSRC and BBSRC for their funding of the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org), BBSRC/EPSRC grant BB/C008219/1, and the Klaus Tschira Foundation (KTF) and the German Federal Ministry of Education and Research (BMBF) for funding the Scientific Databases and Visualization group at the Heidelberg Institute for Theoretical Studies (http://www.h-its.org/). NS also thanks Joseph Dada for assistance with the SBRML export.

References

1 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365–371.

2 Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW *et al.* (2007) The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **25**, 887–893.

3 Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* **26**, 889–896.

4 van Eunen K, Bouwman J, Daran-Lapujade P, Postmus J, Canelas AB, Mensonides FI, Orij R, Tuzun I, van den Brink J, Smits GJ *et al.* (2010) Measuring enzyme activities under standardized *in vivo*-like conditions for systems biology. *FEBS J* **277**, 749–760.

5 Apweiler R, Cornish-Bowden A, Hofmeyr JH, Kettner C, Leyh TS, Schomburg D & Tipton K (2005) The importance of uniformity in reporting protein-function data. *Trends Biochem Sci* **30**, 11–12.

6 Schomburg I, Chang A & Schomburg D (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 47–49.

7 Wittig U, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A, Anstein S, Saric J & Rojas I (2006) SABIO-RK: integration and curation of reaction kinetics data. Proceedings of the 3rd International workshop on Data Integration in the Life Sciences 2006 (DILS'06), Hinxton, UK. *Lect Notes Bioinformatics* **4075**, 94–103.

8 Rojas I, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A & Wittig U (2007) SABIO-RK (System for the Analysis of Biochemical Pathways Reaction Kinetics). *Proceedings of the 2nd International Symposium on "Experimental Standard Conditions of Enzyme Characterizations", 2006, Ruedesheim am Rhein, Germany*, 189–202, Logos-Verlag, Berlin.

9 Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a

community approach to systems biology. *Nat Biotechnol* **26**, 1155–1160.
10 Swainston N & Mendes P (2009) libAnnotationSBML: a library for exploiting SBML annotations. *Bioinformatics* **25**, 2292–2293.
11 Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30.
12 Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M & Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**, D344–D350.
13 The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142–D148.
14 Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S *et al*. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8**, R39.
15 Michaelis L & Menten ML (1913) Die Kinetik der Invertinwirkung. *Biochem Z* **49**, 333–369.
16 Le Novère N (2006) Model storage, exchange and integration. *BMC Neurosci* **7**, S11.
17 Mendes P & Kell DB (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14**, 869–883.
18 Eadie GS (1942) The inhibition of cholinesterase by physostigmine and prostigmine. *J Biol Chem* **146**, 85–93.
19 Hofstee BHJ (1959) Non-inverted versus inverted plots in enzyme kinetics. *Nature* **184**, 1296–1298.
20 Levenberg K (1944) Method for the solution of certain non-linear problems in least squares. *Q Appl Math* **2**, 164–168.
21 Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM J Appl Math* **11**, 431–441.
22 Spasić I, Dunn WB, Velarde G, Tseng A, Jenkins H, Hardy N, Oliver SG & Kell DB (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinform* **7**, 281.
23 Rojas I, Golebiewski M, Kania R, Krebs O, Mir S, Weidemann A & Wittig U (2007) Storing and annotating of kinetic data. *In Silico Biol* **7**(Suppl 2), S37–44.
24 Kell DB. (2006) Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J* **273**, 873–894.
25 Golebiewski M, Mir S, Kania R, Krebs O, Weidemann A, Wittig U & Rojas I (2007) Integration of SABIO-RK in workbenches for kinetic model design. *BMC Syst Biol*, **1**(Suppl 1), P4.
26 Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.
27 Funahashi A, Jouraku A, Matsuoka Y & Kitano H (2007) Integration of CellDesigner and SABIO-RK. *In Silico Biol* **7**(Suppl 2), S81–90.
28 Dada J, Spasić I, Paton N & Mendes P (2010) SBRML: a markup language to associate systems biology data with models. *Bioinformatics* **26**, 932–938.
29 Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P & Oinn T (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* **34**, W729–W732.
30 Li P, Dada JO, Jameson D, Spasić I, Swainston N, Carroll K, Dunn WB, Khan F, Messiha HL, Simeonidis E *et al*. (2010) Systematic integration of experimental data and models in systems biology. *BMC Bioinform* (Under consideration).
31 Swainston N, Jameson D, Li P, Spasić I, Mendes P & Paton NW (2010) Integrative information management for systems biology. *Data Integration in the Life Sciences, Proceedings, 7th International Workshop, DILS 2010* (In press).
32 Vizcaíno JA, Côté R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H & Martens L (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9**, 4276–4283.
33 Anon. (2007) Democratizing proteomics data. *Nat Biotechnol* **25**, 262.
34 Anon. (2007) Time for leadership. *Nat Biotechnol* **25**, 821.
35 Jones P & Côté R (2008) The PRIDE proteomics identifications database: data submission, query, and dataset comparison. *Methods Mol Biol* **484**, 287–303.
36 Siepen JA, Swainston N, Jones AR, Hart SR, Hermjakob H, Jones P & Hubbard SJ (2007) An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ. *Proteome Sci* **5**, 4.
37 Barsnes H, Vizcaíno JA, Eidhammer I & Martens L (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat Biotechnol* **27**, 598–599.
38 Gelperin DM, White MA, Wilkinson ML, Kon Y, Kung LA, Wise KJ, Lopez-Hoyo N, Jiang L, Piccirillo S, Yu H *et al*. (2005) Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev* **19**, 2816–2826.
39 Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK & Weissman JS (2005) Global analysis of protein expression in yeast. *Nature* **425**, 737–741.
40 Malys N & McCarthy JEG (2006) Dcs2, a novel stress-induced modulator of m7GpppX pyrophosphatase

activity that locates to P bodies. *J Mol Biol* **363**, 370–382.

41 Laemmli UK. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T1. *Nature* **227**, 680–685.

42 Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, Fujimoto EK, Goeke NM, Olson BJ & Klenk DC (1985) Measurement of protein using bicinchoninic acid. *Anal Biochem* **150**, 76–85.

Publication 7

An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ.

Siepen JA*, **Swainston N***, Jones AR, Hart SR, Hermjakob H, Jones P, Hubbard SJ.

BMC Proteome Science. 2007, **5**, 4.

**Equal contribution.*

Proteome Science

BioMed Central

Research

Open Access

An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ™

Jennifer A Siepen[†1], Neil Swainston[†2], Andrew R Jones[1,3], Sarah R Hart[4], Henning Hermjakob[5], Philip Jones[5] and Simon J Hubbard*[1]

Address: [1]Faculty of Life Sciences, University of Manchester, M13 9PT, UK, [2]Manchester Interdisciplinary Biocentre, University of Manchester, UK, [3]School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, UK, [4]MBCMS, School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, UK and [5]EMBL Outstation EBI, Wellcome Trust Genome Campus, Hinxton, Cambs, UK

Email: Jennifer A Siepen - jennifer.siepen@manchester.ac.uk; Neil Swainston - neil.swainston@manchester.ac.uk; Andrew R Jones - ajones@cs.man.ac.uk; Sarah R Hart - sarah.hart@manchester.ac.uk; Henning Hermjakob - hhe@ebi.ac.uk; Philip Jones - pjones@ebi.ac.uk; Simon J Hubbard* - simon.hubbard@manchester.ac.uk

* Corresponding author †Equal contributors

Published: 1 February 2007

Proteome Science 2007, **5**:4 doi:10.1186/1477-5956-5-4

Received: 20 December 2006
Accepted: 1 February 2007

This article is available from: http://www.proteomesci.com/content/5/1/4

Abstract

Background: Proteomics continues to play a critical role in post-genomic science as continued advances in mass spectrometry and analytical chemistry support the separation and identification of increasing numbers of peptides and proteins from their characteristic mass spectra. In order to facilitate the sharing of this data, various standard formats have been, and continue to be, developed. Still not fully mature however, these are not yet able to cope with the increasing number of quantitative proteomic technologies that are being developed.

Results: We propose an extension to the PRIDE and mzData XML schema to accommodate the concept of multiple samples per experiment, and in addition, capture the intensities of the iTRAQTM reporter ions in the entry. A simple Java-client has been developed to capture and convert the raw data from common spectral file formats, which also uses a third-party open source tool for the generation of iTRAQTM reported intensities from Mascot output, into a valid PRIDE XML entry.

Conclusion: We describe an extension to the PRIDE and mzData schemas to enable the capture of quantitative data. Currently this is limited to iTRAQTM data but is readily extensible for other quantitative proteomic technologies. Furthermore, a software tool has been developed which enables conversion from various mass spectrum file formats and corresponding Mascot peptide identifications to PRIDE formatted XML. The tool represents a simple approach to preparing quantitative and qualitative data for submission to repositories such as PRIDE, which is necessary to facilitate data deposition and sharing in public domain database. The software is freely available from http://www.mcisb.org/software/PrideWizard.

Background

Proteomics continues to play a critical role in post-genomic science as continued advances in mass spectrometry and analytical chemistry support the separation and identification of increasing numbers of peptides and proteins from their characteristic mass spectra. A desirable trait for such a functional genomics technique is the ability to produce data on a genome-wide basis and, importantly, to be able to do this in a quantitative manner. In proteomics this means being able to quantify the protein changes in different conditions, be they temporal, pathogenic or environmental. Proteomics is beginning to address both these issues; wider genome coverage and quantitation of the proteins present. The latter has been driven by the continued development of techniques for the relative and absolute quantification of protein levels [1-6]. Equally, superior instrumentation and analytical approaches have improved the coverage of genomes, so that genome-wide quantitative proteomics is becoming a reality. This is epitomised by a recent quantitative study acquiring data for the majority of the yeast proteome [7], where the majority of proteins had peptide identifications with available quantitative data obtained using stable isotope labelling in cell culture (SILAC).

Clearly, these types of experiments will become more widespread and detailed. This presents several challenges to the proteomics community and the bioinformatics teams in particular, since it is desirable that this data is captured and stored in appropriate databases in consistent formats, to support data sharing and comparison. Although there are a growing number of data standards [8-12] and databases [13-18] for the storage of proteomic data, at present there is no formal model for quantitative proteomic data that has been fully developed. The Proteome Standards Initiative (PSI) and leading proteomics groups have helped drive the development of several standards for the mass spectral data itself, namely mzXML [8] and mzData [19], and these two are expected to soon merge. These support a comprehensive data model for the storage of proteomic-related mass spectral data, ranging from basic details about the sample, through instrument details and data processing steps, to the actual spectral lists of mass-to-charge values and intensities. This provides a relatively simple yet extensible format for any type of peptide or protein spectra, allowing users to support parent/precursor ion concepts and sophisticated MS^n experiments. Both formats utilise base64 encoding to represent the floating point mass-to-charge (m/z) and ion intensity pairs which form the core of the spectral information. Although this supports the capture of any protein, peptide or fragment ion MS spectra, quantitative data is not explicitly represented in the model. Furthermore, it is not clear how to link the spectra to rich descriptions of the experimental sample, or mixture of samples, within these formats. Indeed, the work to bring this together into a considered whole for proteomics and indeed, in a wider functional genomics context, is well underway, with standards in development for identifications (analysisXML), gels (GelML etc), general sample processing (spML) and functional genomics experiments more generally (FuGE, [11]). Even though the standards development community has not finished this process, database developers in proteomics have already provided solutions for many of these issues in the growing range of proteomics databases now available. These include PeptideAtlas [14], Open Proteomics Database (OPD) [13], Global Proteome Machine (GPM) [15], Pedro [17], PepSeeker [16], and the PRoteomics IDEntifications database (PRIDE) [18,20] as well as others. PeptideAtlas, GPM and PRIDE in particular already contain extensive collections of many millions of peptide identifications. PRIDE, for example, has integrated the mzData data standard into its own PRIDE XML format, which allows users to provide a rich description of their experiment and uses a range of well-supported ontologies to populate the model for a range of meta-data including taxonomy, instrument type, etc. The other databases are also able to capture a similar range of data.

At Manchester, local proteomics groups are active in quantitative proteomics, developing both novel methodology [5,6] and using existing technology to explore quantitative protein levels. In particular, the iTRAQ™ technology [4] is widely used by many groups worldwide, since it offers several advantages, including the ability to multiplex several samples in one single experiment, quantifying several samples in one experiment via a series of reporter ions which are fragmented from an isobaric tag attached to free peptide amines. Thus, researchers can quantify the relative levels of several samples, averaging over data from several peptides, using a labelling technique applicable to all peptides, and not relying on cell culture or similar using stable isotope labelling. This ingenious technique presents informatics with a novel modelling challenge, since such a concept cannot be directly modelled in the existing mzData schema, which considers the sample itself to be a single entity to which all spectra in the experiment are related.

To address this problem, we have conducted a case study to further develop the PRIDE and mzData XML schema to accommodate the concept of multiple samples per experiment, and in addition capture the intensities of the iTRAQ™ reporter ions in the entry. The model extensions are completely compatible with both the PRIDE and mzData schema, utilising controlled vocabulary terms which are added to the respective ontologies. Furthermore, we have developed a simple Java-client (the "Pride Wizard") to capture and convert the raw data from com-

mon spectral file formats, which also uses a third-party open source tool for the generation of iTRAQTM reported intensities from Mascot output. Together, this allows the user to capture large, high-throughput ITRAQTM-based studies, without extensive repetitive manual data entry of individual peptide identifications, and delivered in a valid PSI-consistent data format (PRIDE XML) for submission to the PRIDE repository. The underlying model and Java-client are readily extended to other quantitation techniques. Finally, the Java-client also allows users to directly capture non-quantitative large scale proteomics data, providing the opportunity to convert Mascot-based spectral searches into mzData with associated peptide identifications. We believe this tool will allow proteomics groups to rapidly capture their datasets for submission to a PSI-sanctioned repository and provides a step change in the ease of complex proteomic data available for analysis and sharing for the community in general.

Results

Data capture pipeline

The capture of the mass spectrometry data, associated protein and peptide identifications and quantitative values for multiple samples has been integrated into a single client application, shown in overview in Figure 1. In this example, the mass spectrometry data is represented in Mascot's .mgf format.

As input, the user provides the Pride Wizard with one or more processed mass spectrum files (in either .mgf, mzXML, .pkl or mzData format) and associated Mascot dat files containing protein and peptide identifications. In addition, a number of experimental meta data values are required.

The Pride Wizard can be run in qualitative and quantitative iTRAQTM mode. In the case of the latter, the user specifies a number of samples involved in the experiment, and assigns one or more iTRAQTM labels to each of these samples. Ontology terms can also be assigned to the samples, as the Pride Wizard acts as a client to the EBI Ontology Lookup Service [21] (see Figure 2). A correction factor file must be submitted, in which the isotopic purity of each of the iTRAQTM reagents used in the experiment are specified.

Finally, the user specifies the location of the PRIDE XML file that will be generated upon successful completion of the wizard.

The conversion of this data into valid PRIDE XML takes a number of steps (shown in Figure 1). mzData is required in the final PRIDE XML document, so if necessary, submitted mass spectrum files are converted to mzData using a module provided by ProteomeCommons.org [22]. The ProteomeCommons module is again used to perform conversion of mzData to .mgf files, which are required by the identifications parser module.

The identifications parser takes the form of a Perl script which parses an .mgf mass spectrum file and a Mascot 'dat' file to generate a PRIDE XML fragment containing protein and peptide identifications. In the case where the user has provided iTRAQTM labelled sample data the peptide identification results from the Mascot 'dat' file are merged with iTRAQTM intensities and ratios from i-Tracker [23].

Only the top three ranking identified peptides are reported in the PRIDE XML and the identified peptides are grouped according to the protein accession for the first matching protein for each of the identifications. Where a post-translational modification is assigned by Mascot (fixed or variable) then the name of the modification is matched to the UNIMOD database [24]. If the name of the modification cannot be matched to UNIMOD then the name of the peptide modification is represented as a userParam (see Methods for a description of the schema constructs used).

The i-Tracker software returns the relative ratios of each of the iTRAQTM reporter ions from an .mgf formatted file, a set of correction factors and a user-defined threshold. If the maximum ion peak intensity for any reporter ion peak area is equal to or less than the user-entered threshold a flag of "UT" for "Under Threshold" is reported in the PRIDE XML file.

The iTRAQTM intensities are reported using the iTRAQTM reagent 114 label (see Methods). The actual ratios for each of the iTRAQTM reporter ions, calculated by i-Tracker, are represented as userParams, where, for every peptide identification, we represent the iTRAQTM reporter ion ratios as:

```
<userParam value="1" name="114_114"/>

<userParam value="1.597" name="114_115"/>

...

<userParam value="1.233" name="117_116"/>

<userParam value="1" name="117_117"/>
```

The final step involves merging together each of the mzData files with the generated PRIDE XML fragments to generate a single PRIDE XML document representing the entire experiment. This document is then saved to the user-specified location.

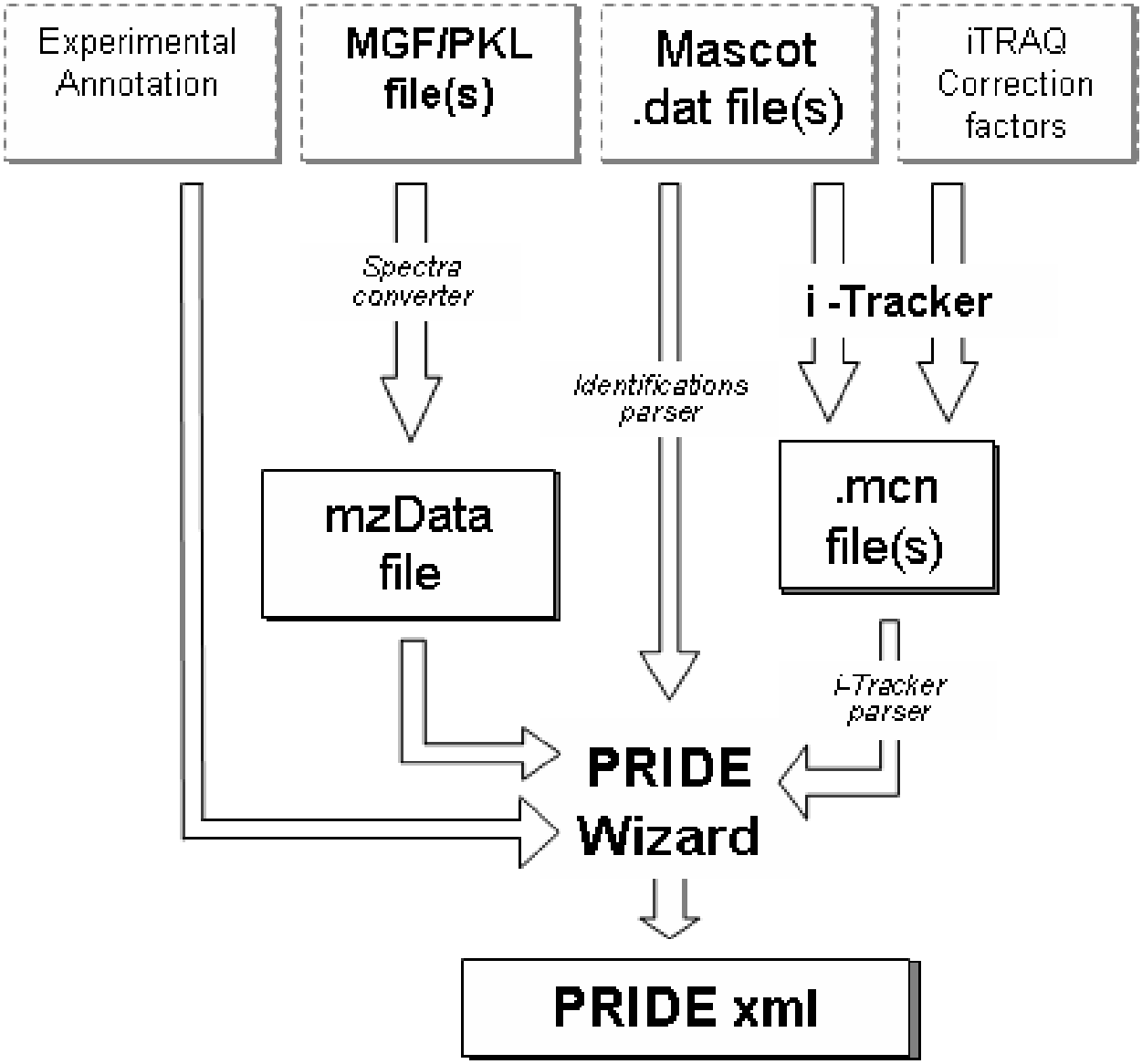

Figure 1
The data processing pipeline for the PRIDE wizard. Grey boxes represent the files/text that are required as input and the black boxes those files created by the PRIDE wizard.

Test data

The software tool was tested on selection of exemplar quantitative data from a number of different collaborating laboratories and successfully created valid PRIDE XML files. The samples included iTRAQTM-based analyses from multiple species, using several instrument types. Full details of the experiments are contained in the methods. The performance of the software was estimated; the wizard takes approximately 4.3 minutes to run on 2314 .mgf formatted mass spectra with 3581 corresponding peptide identifications on a single laptop.

Discussion

We have described a use of controlled vocabulary terms to represent quantitative proteomics data within the PRIDE data format and a software tool to capture and produce the correct file format. Several data standards are currently under development by the Proteomics Standards Initia-

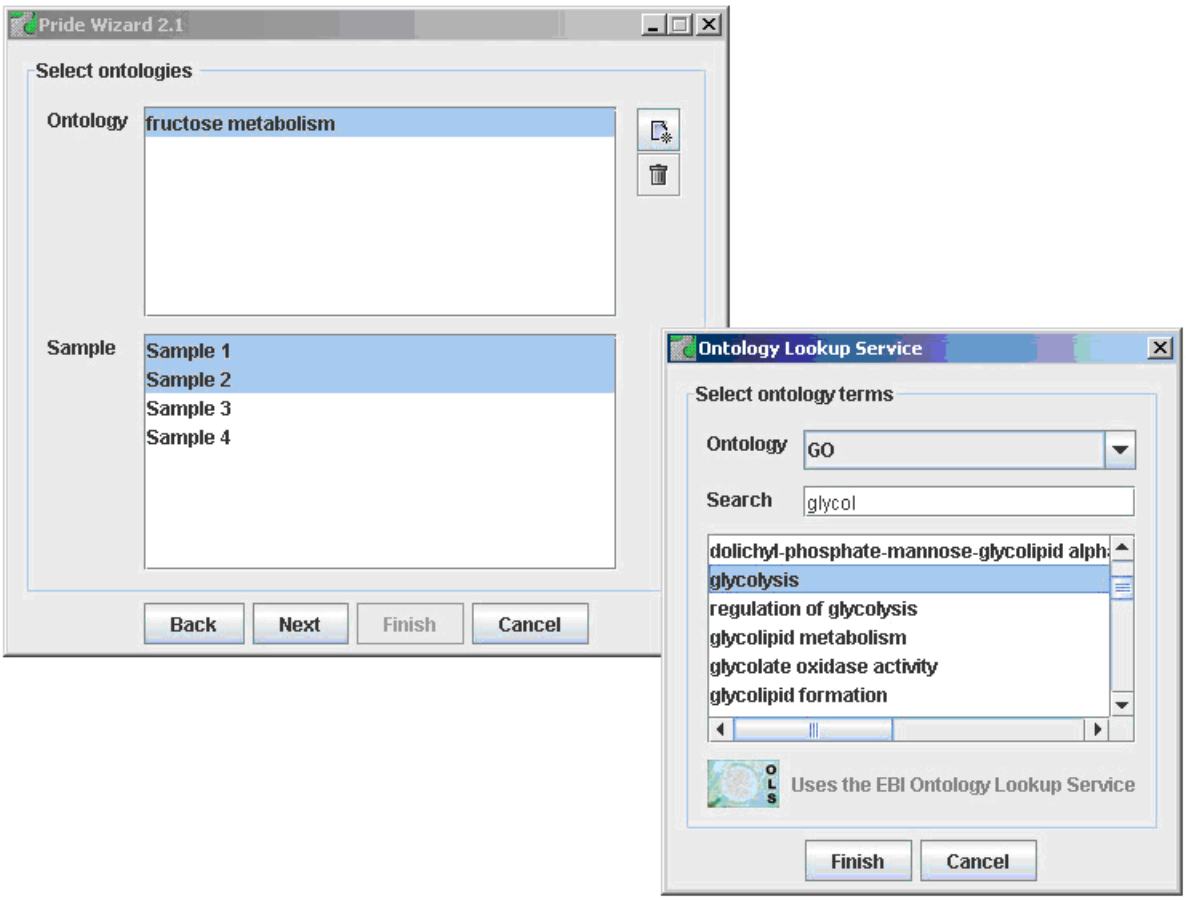

Figure 2
Ontology term selection in the PRIDE Wizard.

tive which will be adopted by PRIDE, allowing a complete proteomics pipeline to be represented. This includes detailed descriptions of protein or peptide separations and labelling (in spML), the mass spectrometry data (mzData) and the protein identifications and quantitative values (analysisXML). However, it is unlikely that spML and analysisXML will be stable and implemented by PRIDE until late 2007 or early 2008. Therefore, the format extension proposed here represents a suitable interim solution for storing quantitative data, and we encourage other laboratories to adopt the conventions. This will allow quantitative data to be represented now in a "pseudo-standard" format and will enable other groups to download such data from PRIDE and perform re-analysis.

In addition to this functionality for iTRAQTM based data, we believe the tool is readily extensible for other quantitative proteomic technologies in a similar fashion, by extending the model and making minor adaptations to the associated Perl and Java code in the Pride Wizard. To this end, we have made the source code available [25]. As data capture needs for both SILAC [1] and QconCAT [5,6] methodologies are underway in our laboratories we expect to provide specific solutions for these approaches in early 2007.

Although the tool was designed ostensibly solely for quantitative data capture, it clearly is able to capture large volumes of identification data and deliver this automatically in PRIDE XML format. We anticipate this will be extremely useful to many groups with high-throughput data sets they wish to capture without tedious manual input. In order to capture the associated experiment, instrument and sample data that can be associated with a PRIDE entry we recommend the PedroDC data capture tool developed at Manchester [26]. Since the PRIDE XML

delivered by our pipeline validates against the PRIDE schema, the data capture tool allows further flexibility to load the PRIDE XML and make suitable additions and edits. Alternatively, the PRIDE team have developed a spreadsheet-based approach linked directly to the Ontology Lookup Service at the EBI which provides an efficient means of entering the higher level data into a PRIDE entry. We anticipate that all of the above will be useful to different user groups, and that a suite of different approaches are probably necessary in proteomics, as any enhancement of data capture capabilities which facilitates data deposition and sharing in public domain repositories is to be welcomed.

Methods

Data capture overview

To generate iTRAQTM quantitative data requires several key component data types which must be integrated. An overview of these data types and the associated analysis tools are shown in Figure 1. A typical iTRAQTM experiment involves the analysis of several samples in a single MS run where peptides are identified in a standard fashion using a search tool such as Mascot [27]. The spectral data are typically delivered to Mascot using Matrix Science's Mascot Generic Format (.mgf), although the tool can cope with a variety of vendor specific formats, as well as mzData. The peptide identifications themselves are contained in Mascot's .dat output file. Finally, to generate the quantitative data for each peptide, users can employ ABI's ProQuant software, or if they prefer, third-party open source tools such as i-Tracker [23]. The latter uses a correction file supplied by ABI to adjust the reporter ion intensities for each identified peptide. The Pride Wizard we have developed integrates these data into a single PRIDE XML file. The model extensions are detailed in the following section.

Modelling quantitative data in PRIDE XML

The mzData schema lies at the heart of a PRIDE XML entry. PRIDE's model is deliberately "light touch" whilst data standards mature, and is readily extensible via inbuilt controlled vocabulary (CV) terms. However, mzData contains only a single sample description object which is also used by PRIDE to capture sample information.

Controlled vocabularies are frequently employed in data formats to provide a consistent extension mechanism allowing a format to capture unanticipated data types [12]. PRIDE files can be annotated with CV and user-defined terms to describe details of the experimental protocol employed, the sample analysed, the instrument used and protein or peptide identifications. We have made use of CV and user-defined terms in PRIDE to support multiplexed sample descriptions and the corresponding quantitative data for each sample (Figure 3)

A CV term in PRIDE has a name (the term itself), a unique accession from the source CV, a label to identify the CV source and optionally a value that can be completed by the user. An example is as follows, where the term *Homo sapiens* (from the NEWT taxonomy) [28] is used to describe the sample:

```
<sampleDescription>

<cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT"/>

</sampleDescription>
```

A further example, where a user-entered value (2000 for "Mass Resolution") has been included:

```
<analyzer>

<cvParam accession="PSI:1000011" name="Mass Resolution" value="2000" cvLabel="PSI"/>

</analyzer>
```

In the extension of PRIDE, we have utilised the userParam facility and the CV representation to capture the names of multiple samples within a single file. The userParam term supports the inclusion of a placeholder for the name of each sub-sample (SUBSAMPLE_1 is used in the example below). This is easily supplemented by the use of additional CV terms in the standard way to add taxonomic or further related information pertaining to the sub-sample. Finally, we have created a further list of CV terms named after the 4 standard iTRAQTM report ions to link the sub-samples to specific reagents called iTRAQTM reagent 114, iTRAQTM reagent 115 etc. The user completes the value attribute of cvParam with the name of each sample they wish to record in the file.

```
<userParam value="Human Liver Extract 1" name="SUBSAMPLE_1"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000114" value="SUBSAMPLE_1" name="iTRAQ reagent 114">
```

The term SUBSAMPLE_1 then serves as a unique identifier for that sample throughout the rest of the file. Where the user wants to add further CV terms to describe the sample, the value attribute is completed with SUBSAMPLE_1.

```
<sampleDescription>

<cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT" value="SUBSAMPLE_1"/>
```

Standard PRIDE entry

```
<Experiment>
  <Title>Human liver proteome investigation</Title>
  <ShortLabel>HumanLiver</ShortLabel>
  <Protocol>
    <ProtocolName>LC-MS analysis of liver proteome</ProtocolName>
  </Protocol>
  <mzData version="1.05" accessionNumber="1234">
    <description>
      <admin>
        <sampleName>Human liver extract 1</sampleName>
        <sampleDescription>
          <cvParam name="Homo sapiens" accession="NEWT:9606" cvLabel="NEWT"/>
          <cvParam name="liver" accession="MeSH:D008099" cvLabel="MeSH"/>
        </sampleDescription>
      .....
```

Single sample description

```
<GelFreeIdentification>
  <Accession>NP_000537</Accession>
  <Database>Genbank</Database>
  <PeptideItem>
    <Sequence>meepqsdpsvepplsdpgp</Sequence>
    <SpectrumReference>1</SpectrumReference>
  </PeptideItem>
  <PeptideItem>
    <Sequence>lpennvlsplpsqamddlmlspddieqwfte</Sequence>
    <SpectrumReference>2</SpectrumReference>
  </PeptideItem>
  <Score>254</Score>
  <SearchEngine>MASCOT</SearchEngine>
</GelFreeIdentification>
.....
```

Identification details corresponds with the sample entered

Extended PRIDE entry

```
<Experiment>
  <Title>Human liver/kidney proteome investigation</Title>
  <ShortLabel>HumanLiver</ShortLabel>
  <Protocol>
    <ProtocolName>LC-MS iTRAQ analysis of liver versus kidney proteome</ProtocolName>
  </Protocol>
  <mzData version="1.05" accessionNumber="1235">
    <description>
      <admin>
        <sampleName>Mixed sample for iTRAQ analysis</sampleName>
        <sampleDescription >
          <userParam value="Human liver extract 1" name="SUBSAMPLE_1" />
          <cvParam cvLabel="NEWT" accession="NEWT:9606" name="Homo sapiens" value="SUBSAMPLE_1" />
          <cvParam cvLabel="MeSH" accession="MeSH:D008099" name="liver" value="SUBSAMPLE_1" />
          <cvParam cvLabel="PRIDE" accession="PRIDE:0000114 name="iTRAQ reagent 114" value="SUBSAMPLE_1" />

          .....

          <userParam value="Human kidney extract 1" name="SUBSAMPLE_3" />
          <cvParam cvLabel="NEWT" accession="NEWT:9606" name="Homo sapiens" value="SUBSAMPLE_3" />
          <cvParam cvLabel="MeSH" accession="MeSH:D007668" name="kidney" value="SUBSAMPLE_3" />
          <cvParam cvLabel="PRIDE" accession="PRIDE:0000116" name="iTRAQ reagent 116" value="SUBSAMPLE_3" />
          ....
        </sampleDescription>
    .....
```

Unique identifier used to assign CV terms describing each sample

Unique identifier assigned to each sample

```
<GelFreeIdentification>
  <Accession>NP_000537</Accession>
  <Database>Genbank</Database>

  <PeptideItem>
    <Sequence>meepqsdpsvepplsdpgp</Sequence>
    <SpectrumReference>1</SpectrumReference>
    <additional>
      <cvParam cvLabel="PRIDE" accession="PRIDE:0000118" value="345.0334" name="iTRAQ intensity 114"/>
      <cvParam cvLabel="PRIDE" accession="PRIDE:0000119" value="9450.0034" name="iTRAQ intensity 115"/>
      <cvParam cvLabel="PRIDE" accession="PRIDE:0000120" value="132.45" name="iTRAQ intensity116"/>
      <cvParam cvLabel="PRIDE" accession="PRIDE:0000121" value="890.1" name="iTRAQ intensity 117"/>
    </additional>
  </PeptideItem>
  ....
</GelFreeIdentification>
.....
```

Intensity values measured for each peak, corresponding to specific iTRAQ reagents

Figure 3
Extensions to the PRIDE XML schema.

</sampleDescription>

In order to enter the actual intensities of the reporter ions, we propose the following convention, adapting the iTRAQTM reagent 114 label further to iTRAQTM intensity 114 as shown below.

<cvParam cvLabel="PRIDE" accession="PRIDE:0000118" value="0.048" name="iTRAQ intensity 114"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000119" value="0.193" name="iTRAQ intensity 115"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000120" value="0.204" name="iTRAQ intensity 116"/>

<cvParam cvLabel="PRIDE" accession="PRIDE:0000121" value="0.65" name="iTRAQ intensity 117"/>

The terms proposed here have been added to the PRIDE CV and have been assigned stable accession numbers. It is apparent from this example that other quantitative data with intensity or ratio values, calculated in a variety of ways, can be represented using similar CV terms.

Test data

The software tool was tested on a number of different data sets from different laboratories. Test set 1 was derived from *Trypanosoma brucei* flagellum samples which were prepared as described previously [29]. Samples were derivatised using iTRAQTM according to the manufacturer's instructions and derivatised peptides from four samples were prepared and analysed online with a QTOF I instrument (Waters, Manchester, upgraded to QTOF II specifications by MS Horizons, Manchester). Data acquisition was performed using MassLynx 3.4, acquiring 3 channels of tandem MS data. Following acquisition, data were processed using ProteinLynx to generate .pkl files.

Test set 2 was derived from soluble extracts from the gram negative plant pathogenic bacterium *Erwinia carotovora* (sp atroseptica SCRI1043) which were prepared as described previously [30]. Three biological replicate samples were labelled with iTRAQTM reagents 114–116 respectively, a fourth sample which consisted of a pool of the three replicates was labelled with the 117 iTRAQTM reagent. Labelling, multidimensional LC and MSMS were carried out as in [31]. The data submitted to the PRIDE wizard was essentially from the combination of running four fractions from strong cation exchange column on LCMSMS (QSTAR, Applied Biosystems).

Test set 3 was derived from primitive hematopoietic cells from mouse bone marrow as described previously [32]. Samples were derivatised using iTRAQTM according to the manufacturer's instructions and derivatised peptides from four samples were prepared and analysed online with a QSTAR XL (Applied Biosystems). Data acquisition was performed using an independent data acquisition protocol as described previously [32].

Availability and requirements

Project name: Pride Wizard

Project homepage: http://www.mcisb.org/software/PrideWizard

Operating system: Windows

Programming language: perl, Java 1.4.2 and above.

Licence: GNU GPL

Authors' contributions

SJH, ARJ, JAS, NS & SRH designed and developed the quantitative data model in the PRIDE XML schema, and jointly proposed the schema extensions with subsequent additional verification from PJ & HH. JAS & NS developed the software tool to capture the data, with testing from SRH. SJH conceived the study, lead the manuscript production with contributions from all authors, who have read and approved the final manuscript.

Declaration of competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

The authors are indebted to Prof Tony Whetton and Dr Richard Unwin at the University of Manchester, Dr Kathryn Lilley at Cambridge University for provision of data sets to evaluate the wizard, and Jayson Falkner of the University of Michigan for assistance with the ProteomeCommons module. This work has been supported by several BBSRC and EPSRC grants to the authors, ISPIDER (BBSB17204) to JAS and SJH, Pedro (BBSB12407) to ARJ, SJH, and the Manchester Centre for Integrative Systems Biology which supported NS.

References

1. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** *Molecular & Cellular Proteomics* 2002, **1(5):**376-386.
2. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP: **Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(12):**6940-6945.
3. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nature Biotechnology* 1999, **17(10):**994-999.
4. Ross PL, Huang YLN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.** *Molecular & Cellular Proteomics* 2004, **3(12):**1154-1169.

5. Beynon RJ, Doherty MK, Pratt JM, Gaskell SJ: **Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides.** *Nature Methods* 2005, **2(8):**587-589.
6. Pratt JM, Simpson D, Doherty M, Rivers J, Gaskell SJ, Beynon RJ: **Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes.** *Nature Protocols* 2006, **1:**1029-1043.
7. de Godoy LMF, Olsen JV, de Souza GA, Li GQ, Mortensen P, Mann M: **Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system.** *Genome Biology* 2006, **7(6):**.
8. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu WM, Aebersold R: **A common open representation of mass spectrometry data and its application to proteomics research.** *Nature Biotechnology* 2004, **22(11):**1459-1466.
9. Taylor CF, Hermjakob H, Julian RK, Garavelli JS, Aebersold R, Apweiler R: **The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI).** *Omics-a Journal of Integrative Biology* 2006, **10(2):**145-151.
10. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik R, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li YX, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu WM, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios L, Eisenberg D, Steipe B, Hogue C, Apweiler R: **The HUPOPSI's Molecular Interaction format - a community standard for the representation of protein interaction data.** *Nature Biotechnology* 2004, **22(2):**177-183.
11. Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, DeGreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK, Laursen K, Oliver SG, Paton NW, Sarkans U, Sansone S, Stoeckert CJ, Taylor CF, Whetzel PL, White JA, Spellman P, Piazarro A: **The functional genomics experimental model (FuGE): an extensible framework for standards in functional genomics.** *Nature Biotechnology* 2006, **in press.:**.
12. Jones AR, Paton NW: **An analysis of extensible modelling for functional genomics data.** *Bmc Bioinformatics* 2005, **6:**.
13. Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM: **The need for a public proteomics repository.** *Nature Biotechnology* 2004, **22(4):**471-472.
14. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Research* 2006, **34:**D655-D658.
15. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *Journal of Proteome Research* 2004, **3(6):**1234-1242.
16. McLaughlin T, Siepen JA, Selley J, Lynch JA, Lau KW, Yin HJ, Gaskell SJ, Hubbard SJ: **PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns.** *Nucleic Acids Research* 2006, **34:**D649-D654.
17. Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, Stead D, Yin Z, Brown AJP, Hesketh A, Chater K, Hansson L, Mewissen M, Ghazal P, Howard J, Lilley KS, Gaskell SJ, Brass A, Hubbard SJ, Oliver SG, Paton NW: **PEDRo: A database for storing, searching and disseminating experimental proteomics data.** *Bmc Genomics* 2004, **5:**.
18. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R: **PRIDE: The proteomics identifications database.** *Proteomics* 2005, **5(13):**3537-3545.
19. Orchard S, Hermjakob H, Taylor CF, Potthost F, Jones P, Zhu WM, Julian RK, Apweiler R: **Second Proteomics Standards Initiative Spring Workshop.** *Expert Review of Proteomics* 2005, **2(3):**287-289.
20. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community.** *Nucleic Acids Research* 2006, **34:**D659-D663.
21. Cote RG, Jones P, Apweiler R, Hermjakob H: **The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.** *Bmc Bioinformatics* 2006, **7:**.
22. Falkner JA, Ulintz PJ, Andrews PC: **http://www.proteomecommons.org/.** .
23. Shadforth IP, Dunkley TPJ, Lilley KS, Bessant C: **i-Tracker: For quantitative proteomics using iTRAQ (TM).** *Bmc Genomics* 2005, **6:**145.
24. Creasy DM, Cottrell JS: **Unimod: Protein modifications for mass spectrometry.** *Proteomics* 2004, **4(6):**1534-1536.
25. **PrideWizard** [http://www.mcisb.org/software/PrideWizard]
26. **PedroDC Capture Tool** [http://pedrodownload.man.ac.uk]
27. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20(18):**3551-3567.
28. Phan IQH, Pilbout SF, Fleischmann W, Bairoch A: **NEWT, a new taxonomy portal.** *Nucleic Acids Research* 2003, **31(13):**3822-3823.
29. Broadhead R, Dawe HR, Farr H, Griffiths S, Hart SR, Portman N, Shaw MK, Ginger ML, Gaskell SJ, McKean PG, Gull K: **Flagellar motility is required for the viability of the bloodstream trypanosome.** *Nature* 2006, **440(7081):**224-227.
30. Coulthurst SJ, Lilley KS, Salmond GPC: **Genetic and proteomic analysis of the role of luxS in the enteric phytopathogen, Erwinia carotovora.** *Molecular Plant Pathology* 2006, **7(1):**31-45.
31. Dunkley TPJ, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, Griffin JL, Bessant C, Brandizzi F, Hawes C, Watson RB, Dupree P, Lilley KS: **Mapping the Arabidopsis organelle proteome.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(17):**6518-6523.
32. Unwin RD, Smith DL, Blinco D, Wilson CL, Miller CJ, Evans CA, Jaworska E, Baldwin SA, Barnes K, Pierce A, Spooncer E, Whetton AD: **Quantitative proteomics reveals posttranslational control as a regulatory factor in primary hematopoietic stem cells.** *Blood* 2006, **107(12):**4687-4694.

Publication 8

A QconCAT informatics pipeline for the analysis, visualisation and sharing of absolute quantitative proteomics data.

Swainston N, Jameson D, Carroll K.

Proteomics. 2011, **11**, 329–33.

TECHNICAL BRIEF

A QconCAT informatics pipeline for the analysis, visualization and sharing of absolute quantitative proteomics data

Neil Swainston[1,2], *Daniel Jameson*[1,2] *and Kathleen Carroll*[1,3]

[1] Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK
[2] School of Computer Science, University of Manchester, Manchester, UK
[3] School of Chemistry, University of Manchester, Manchester, UK

Absolute protein concentration determination is becoming increasingly important in a number of fields including diagnostics, biomarker discovery and systems biology modeling. The recently introduced quantification concatamer methodology provides a novel approach to performing such determinations, and it has been applied to both microbial and mammalian systems. While a number of software tools exist for performing analyses of quantitative data generated by related methodologies such as SILAC, there is currently no analysis package dedicated to the quantification concatamer approach. Furthermore, most tools that are currently available in the field of quantitative proteomics do not manage storage and dissemination of such data sets.

Received: July 21, 2010
Revised: September 29, 2010
Accepted: October 18, 2010

Keywords:
Bioinformatics / Data analysis / Data management / Quantification concatamer / Quantitation

An informatics workflow is introduced, which represents a solution to the data analysis and management challenges encountered in applying the quantification concatamer (QconCAT) methodology to absolute quantitative proteomics studies. The workflow includes automated database searching and quantitation, and data storage and sharing, utilizing existing tools and applying community-developed standards where possible. A study of glycolytic enzymes from *Saccharomyces cerevisiae* is discussed, alongside a comparison of absolute protein concentrations calculated by these tools and by manual analysis

Absolute protein concentration is becoming increasingly important in a number of fields including systems biology modeling. The QconCAT methodology [1] is a recently introduced approach for determining absolute protein concentration. It involves expressing concatenated constructs of reporter peptides in *Escherichia coli* and growing these in culture media supplemented with isotopically labeled amino acid analogues. This forms a recombinant QconCAT protein, containing isotopically labeled peptides that act as unique internal standards for each of the proteins of interest. Known concentrations of this labeled protein can then be introduced to a given sample, which upon digestion yield equimolar amounts of QconCAT peptides, and when co-digested with endogenous proteins will produce pairs of heavy labeled and light native peptides, with identical chemical properties. Following analysis by LC-MS, the co-elution of such peptide pairs allows their relative abundance to be calculated from the response ratio between the analyte and the internal standard peptide. Absolute concentrations of each target protein can then be inferred from the known concentrations of labeled analyte.

The informatics steps involved in performing such an experiment are as follows: (i) selection of suitable reporter

Correspondence: Neil Swainston, Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK
E-mail: neil.swainston@manchester.ac.uk
Fax: +44-161-306-8918

Abbreviation: QconCAT, quantification concatamer

Colour Online: See the article online to view figs. 1 and 2 in colour.

peptide(s) for each protein to be quantified; (ii) data acquisition and identification of peptides; (iii) quantitation of peptide pairs and inference of absolute protein concentrations; and (iv) storage and dissemination of identifications, quantitations and associated mass spectra. There is currently no dedicated informatics support for steps (ii)–(iv). This pipeline rectifies this by providing an integrated system consisting of two parts: an analysis tool to perform peptide identification and quantitation, and a database repository allowing this data to be stored and visualized (see Fig. 1).

Analysis is performed by the QconCAT PrideWizard, an extension of the original PrideWizard [2], which was developed to quantify iTRAQ labeled samples. The wizard provides a user interface to which batches of spectra may be submitted. Labeled peptides are then identified through a Mascot [3] MS/MS Ion Search. Protein hits are filtered such that only those that contain at least one QconCAT labeled peptide, with rank 1 and a peptide expect score <5, are retained. Furthermore, peptides are filtered such that the only ones quantified are unmodified (apart from the QconCAT label) and are unique to a single protein.

Quantitation is then performed by firstly generating an extracted ion chromatogram for the m/z value corresponding to the precursor ion matching each labeled peptide. Where multiple matches occur against the same labeled peptide in a given protein, the highest scoring one is considered for quantitation. Savitzky-Golay smoothing [4] is applied to the chromatogram, and the start and end retention time for the chromatographic peak matching the peptide is determined, based on the retention time of the fragmentation spectrum that supplied the peptide match. Each precursor scan within this retention time window is extracted and analyzed with an implementation of the SILAC Analyzer linear fit quantitation algorithm [5]. This provides a light/heavy ratio, and standard error, for each identified unlabeled/labeled peptide pair (see Fig. 2). Absolute quantification is obtained by multiplying this ratio by the known amount of the standard. Protein quantitations are a function of individual peptide quantitations and are calculated using a formula described previously [6]. Peptides from each replicate contribute to the overall protein quantitation.

Upon completion of the analysis, raw experimental data, metadata, protein and peptide identifications and quantitations are formatted into PRIDE XML [7] (see Supporting Information Fig. S1). The PRIDE XML documents are automatically uploaded to a native XML database, where they can be queried through both a web and web service interface (see http://www.mcisb.org/QconCAT/). The web interface provides a simple search facility, allowing proteins to be searched by UniProt identifiers. Protein summaries display the calculated ratio, labeled peptides from each of the submitted replicates, a link to the original Mascot results, a tooltip showing experimental metadata and an interactive panel, allowing fragmentation spectra, precursor spectra and extracted mass chromatograms to be viewed. In this way, both identifications and quantitations can be viewed and shared online, displaying the original raw data as was acquired by the instrument (see Fig. 2). Data can be copied from the web interface and pasted into spreadsheets or e-mails, providing an export facility for report generation. The system can also be configured to manage the original, vendor-specific raw data file, which can then also be downloaded from the web interface.

To test the system, a study was performed upon yeast glycolytic enzymes with the results generated by the QconCAT PrideWizard compared to those generated manually. Samples were prepared and data collected in triplicate by LC-MS using a nanoACQUITY chromatograph (Waters MS Technologies, Manchester, UK) coupled to an LTQ-Orbitrap

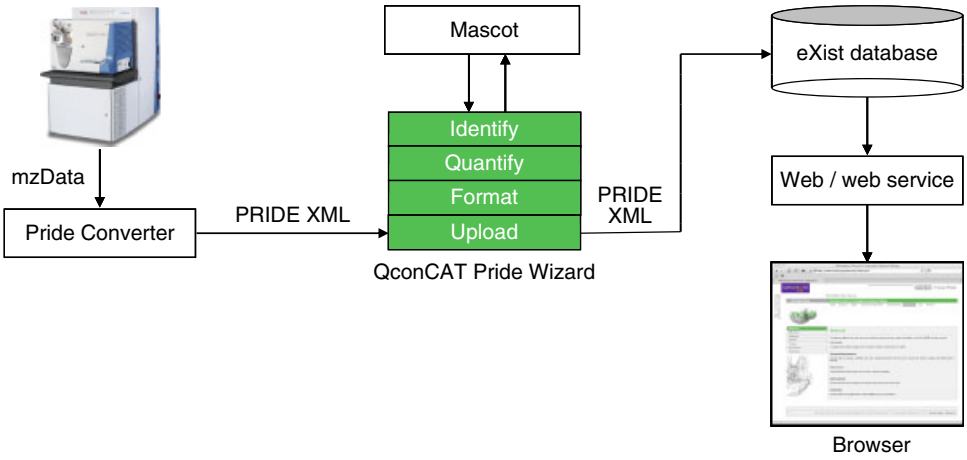

Figure 1. The QconCAT analysis pipeline. Data are exported from the instrument software in mzData format, which is then imported into the PRIDE Converter where metadata is applied. Batches of resulting PRIDE XML may be input to the QconCAT PrideWizard, which submits queries to Mascot and quantifies the resulting identified peptides. Identifications and quantitations are merged with the spectral data, and the resulting document is uploaded to a native XML database where it can be accessed through a web and web service interface.

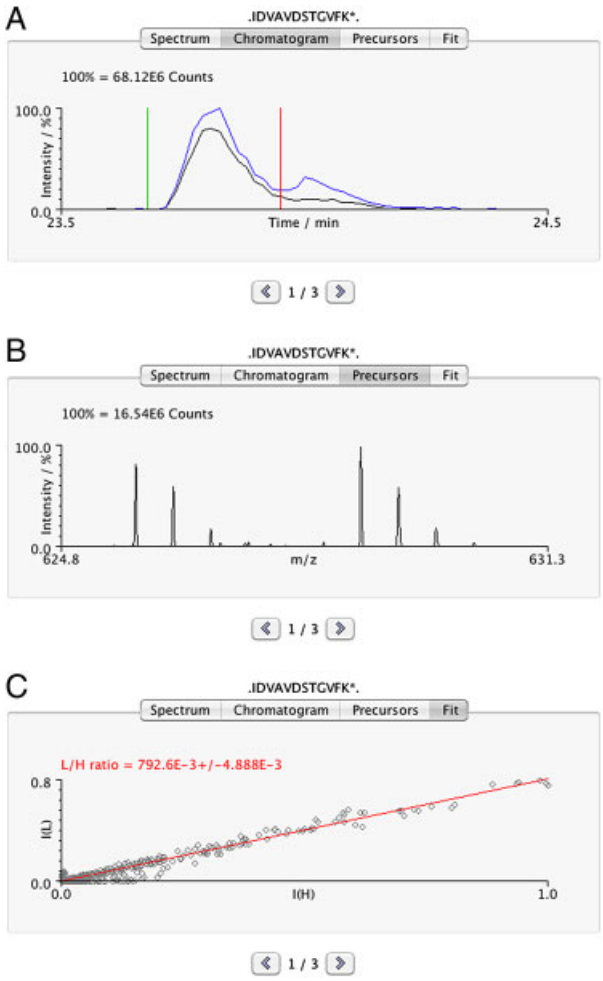

Figure 2. Web interface screen captures providing visualization of the quantitation of the peptide IDVAVDSTGVFK from G3P1_YEAST. This represents the first of three labeled peptides that were assigned to this protein (see http://www.mcisb.org/QconCAT/G3P1_YEAST/). (A) shows the extracted ion chromatograms for the masses 625.836 and 628.846 Da, representing the unlabeled and labeled peptide in black and blue, respectively, and illustrates their co-elution. Vertical lines indicate the calculated start and end retention times of the labeled peptide chromatographic peak. (B) illustrates a precursor scan taken between the start and end retention times highlighted in (A), containing isotopic envelopes for both unlabeled and labeled peptides. The SILAC Analyzer linear fit algorithm is applied to each of these precursor scans. This entails applying a sliding window across the isotopic clusters in each of the precursor scans (B), gathering pairs of intensity readings at m/z (representing the unlabeled peptide) and $m/z+\Delta m/z$ (representing the labeled peptide, where Δm is the monoisotopic mass of the label (6.020 Da in these studies) and z is the precursor ion charge). If both the labeled and unlabeled peptides are present, these intensity pairs display a linear correlation, which is plotted in the Fit tab (C). Applying linear regression to this scatter plot provides an unlabeled to label intensity ratio and standard error, which in the above example is 0.793 ± 0.005.

(ThermoFisher Scientific, Waltham, MA, USA). The acquired raw data were converted to the vendor-independent mzData [8] format using Bioworks Browser (v3.3.1 SP3, ThermoFisher, Bremen, Germany), an operation that performs no operations such as deisotoping or charge deconvolution. These data were passed through the PRIDE Converter [9] version 2.2, metadata was added and the data exported in PRIDE XML format. Analysis was then performed on all replicates using the QconCAT PrideWizard. Manual analysis was also performed through generation of peak areas from extracted ion chromatograms (Bioworks Browser). Ratios of unlabeled to labeled peptide areas were calculated and these averaged across replicates.

Results show that the QconCAT PrideWizard is able to quantify 23 of the 27 proteins in the study, in comparison to 19 by manual analysis (see Table 1). Correlation is observed between the 17 proteins quantified by both the QconCAT PrideWizard and by manual analysis ($R^2 = 0.88$). The QconCAT PrideWizard additionally quantified five low abundance proteins, with the least abundant ADH5_YEAST reported at 10^4 copies per cell. ALF_YEAST was successfully quantified manually, but could not be quantified by the QconCAT PrideWizard due to neither of the two labeled marker peptides representing this protein being found by

Table 1. Protein concentrations calculated by the QconCAT PrideWizard and by manual analysis.

Protein accession	Protein concentration/molecules per cell	
	QconCAT PrideWizard	Manual analysis
ADH1_YEAST	9.718E+04	1.492E+05
ADH2_YEAST	2.001E+04	
ADH3_YEAST	7.518E+04	
ADH4_YEAST	5.693E+05	5.692E+05
ADH5_YEAST	1.049E+04	
ADH6_YEAST		
ADH7_YEAST		
ALF_YEAST		5.032E+06
ENO1_YEAST	3.348E+06	3.552E+06
ENO2_YEAST	8.998E+06	1.035E+07
G3P1_YEAST	1.799E+06	1.766E+06
G3P2_YEAST		1.623E+07
G3P3_YEAST	1.822E+07	1.623E+07
G6PI_YEAST	5.488E+05	5.646E+05
HXKA_YEAST	7.157E+04	8.513E+04
HXKB_YEAST	2.499E+05	2.490E+05
HXKG_YEAST	1.192E+05	1.430E+05
K6PF1_YEAST	1.806E+05	1.728E+05
K6PF2_YEAST	1.499E+05	1.702E+05
KPYK1_YEAST	9.535E+06	1.523E+07
KPYK2_YEAST	2.032E+04	
PDC1_YEAST	4.620E+06	4.869E+06
PDC5_YEAST	4.343E+04	
PDC6_YEAST	1.685E+04	
PGK_YEAST	1.103E+06	1.518E+06
PMG1_YEAST	2.958E+06	3.234E+06
TPIS_YEAST	1.020E+06	2.394E+06

Mascot. Conversely, reporter peptides for ADH7_YEAST were identified, but the corresponding native peptides were in such low abundance that the protein could be quantified. No peptides were identified for ADH6_YEAST.

The study attempts to quantify a number of isoenzymes, including two members of the enolase family. A number of QconCAT peptides originally selected to act as unique marker peptides for a given protein were found to be duplicates; that is, they were shared between a number of isoenzymes. An example is SGETEDTFIADLVVGLR, originally selected as a marker peptide for ENO1_YEAST, which was correctly ignored in the quantification calculation of the QconCAT PrideWizard on the grounds that it is also present in ENO2_YEAST. As it is common practice to select multiple peptides to act as markers for a given protein, ENO1_YEAST was successfully quantified by the unique peptides TFAEALR and NVNDVIAPAFVK. However, G3P2_YEAST was not quantified, as both selected marker peptides (VLPELQGK and VPTVDVSVVDLTVK) were non-unique. Such common peptides may be considered in unbiased approaches such as SILAC analyses, where the contribution that a shared peptide makes to each of its proteins may be inferred. However, it is appropriate to exclude such peptides from QconCAT studies, as their presence can usually be mitigated by experimental design.

The approach taken by the QconCAT Pride Wizard is to first identify heavy/light peptide pairs and then to quantify them. The identification of QconCAT pairs is driven by the Mascot MS/MS Ion Search matching labeled QconCAT peptides, which is dependent upon acquisition of MS/MS data for these peptides. While this may not be applicable to relative quantitative proteomics studies such as SILAC, where a sample may contain hundreds or thousands of pairs across a large dynamic range, QconCAT studies focus on a finite number of peptide pairs (typically ~50). Furthermore, it is assumed the labeled QconCAT protein is added to the sample in a concentration large enough to ensure that its peptides provide signals of sufficient intensity that fragmentation data will be acquired. Quantitation is performed by extracting individual survey scans containing each heavy/light peptide pair and determining signal ratios of equivalent points across the two isotopic clusters. Such an approach minimizes the potentially detrimental effect of co-eluting peptides.

A key feature of the system is its performance. A single raw data file of 133 MB can be quantified in 249 s (of which 132 s are accounted for by the Mascot database search itself) on a MacBook Pro (Apple Corporation, CA, USA), running Mac OS X 10.6.4, with a 2.5-GHz Intel Core 2 Duo processor and 4 GB 667 MHz DDR2 SDRAM.

A further consideration in the development of the system was to ensure ease-of-use, as the system has been designed for use by mass spectrometrists rather than bioinformaticians. The QconCAT PrideWizard manages the flow of data from spectral data submission through database searching, peptide/protein quantitation, data formatting and storage. As such, a user can submit a batch of spectral data files, which with only a small number of additional inputs allows all steps of the analysis pipeline to be performed in a single operation.

Existing tools such as Mascot and the SILAC Analyzer have been reused and repackaged: a deliberate strategy that avoids wheel reinvention. Where possible, the workflow has utilized existing data representation standards. Mascot has been recognized as the de facto industry standard for performing protein identification and a link to the original Mascot results, along with individual peptide scores, is provided. Reanalysis of the original raw data is possible due to the facility to export the data in standard, non-proprietary mzData format. Peptide and protein identifications, matched spectra and experimental metadata can be viewed, queried and exported. This corresponds to journal recommendations, which state that original, raw experimental data should be made available, along with the secondary derived data in terms of protein identifications and quantitations in a standardized format that facilitates query and use by third parties [10]. Furthermore, the authors intend to update the QconCAT pipeline to support subsequent iterations of the PRIDE XML format that incorporate the newly introduced HUPO Proteomics Standards Initiative standards mzML, mzIdentML and ultimately, mzQuantML.

The QconCAT Browser allows the survey scan data upon which quantitations are performed to be viewed, allowing both identifications and quantitations to be verified by visualization of the original fragmentation and survey spectra in a web browser. This contrasts with the usual procedure, in which verification of reported quantitative data is rarely achievable due to the inaccessibility of the original raw data. Even when raw data are accessible, it is commonly held upon the instrument computer and can usually only be accessed through the vendor-supplied instrumentation software. Explicitly displaying quantitative data along with peptide and protein identifications in a web accessible manner provides significant benefits and is an approach that will hopefully become more widespread over time.

In addition to the web browser interface, a web service interface is provided, allowing programmatic access to the concentration values, identifications and all spectra contained in the database. The web service interface allows the user to submit customized XQuery commands to the XML database, providing the flexibility to query and retrieve any element of the PRIDE XML document, from individual peptide or protein records up to the entire document itself.

This QconCAT analysis pipeline provides a freely available, vendor-independent means of analyzing, visualizing and disseminating QconCAT experimental data, performing the calculation of absolute protein concentrations and managing the storage and dissemination of these data in a standards compliant manner.

The authors thank Kieran Smallbone, Norman Paton, Rob Beynon, Claire Eyers, Simon Hubbard, Craig Lawless and Julian Selley, and Marcin Rzeznicki for his Savitzky-Golay

algorithm implementation (http://code.google.com/p/savitzky-golay-filter/). The authors thank the EPSRC and BBSRC for their funding of the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org), BBSRC/EPSRC Grant BB/C008219/1.

The authors have declared no conflict of interest.

References

[1] Beynon, R. J., Doherty, M. K., Pratt, J. M., Gaskell, S. J., Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* 2005, *2*, 587–589.

[2] Siepen, J. A., Swainston, N., Jones, A. R., Hart, S. R. *et al.*, An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQ. *Proteome Sci.* 2007, *5*, 4.

[3] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[4] Savitzky, A., Golay, M. J. E., Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 1964, *36*, 1627–1639.

[5] Nilse, L., Sturm, M., Trudgian, D. *et al.*, SILACAnalyzer – a tool for differential quantitation of stable isotope derived data. *CIBB, 6th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, Genoa 2009.

[6] Baker, R. W. R., Nissim, J. A., Expressions for combining standard errors of two groups and for sequential standard error. *Nature* 1963, *198*, 1020.

[7] Jones, P., Côté, R. G., Martens, L., Quinn, A. F. *et al.*, PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 2006, *34*, D659–D663.

[8] Orchard, S., Taylor, C., Hermjakob, H., Zhu, W. *et al.*, Current status of proteomic standards development. *Expert Rev. Proteomics* 2004, *1*, 179–183.

[9] Barsnes, H., Vizcaíno, J. A., Eidhammer, I., Martens, L., PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.* 2009, *27*, 598–599.

[10] Anon. Democratizing proteomics data. *Nat. Biotechnol.* 2007, *25*, 262.

Publication 9

Integrative Information Management for Systems Biology.

Swainston N*, Jameson D*, Li P*, Spasić I, Mendes P, Paton NW.

In proceedings of the 7th International workshop on Data Integration in the Life Sciences 2010 (DILS'10), Gothenburg, Sweden. Lecture Notes in Computer Science. 2010, **6254**, 164-178.

**Equal contribution.*

Integrative Information Management for Systems Biology

Neil Swainston, Daniel Jameson, Peter Li, Irena Spasic, Pedro Mendes, and Norman W. Paton

School of Computer Science, University of Manchester,
Oxford Road, Manchester M13 9PL, UK
{neil.swainston,daniel.jameson,peter.li,i.spasic,
pedro.mendes,npaton}@manchester.ac.uk

Abstract. Systems biology develops mathematical models of biological systems that seek to explain, or better still predict, how the system behaves. In bottom-up systems biology, systematic quantitative experimentation is carried out to obtain the data required to parameterize models, which can then be analyzed and simulated. This paper describes an approach to integrated information management that supports bottom-up systems biology, with a view to automating, or at least minimizing the manual effort required during, creation of quantitative models from qualitative models and experimental data. Automating the process makes model construction more systematic, supports good practice at all stages in the pipeline, and allows timely integration of high throughput experimental results into models.

Keywords: computational systems biology, workflow.

1 Introduction and Motivation

Systems biology involves the development and study of mathematical models of biological systems. Existing databases of pathways, combined with the emergence of consensus models of specific organisms [1], provide broad access to qualitative models of biological systems. In bottom-up systems biology, these qualitative models can be used as a starting point for the creation of quantitative models that support a range of different forms of simulation and analysis [2]. As such, bottom-up systems biology projects: (i) identify the pathway or portion of a network that is to be modeled; (ii) associate the model with functions and parameter values that represent its dynamic behavior, either from databases [3] or experimentation; and (iii) analyze and/or simulate the resulting model to understand its properties.

In common practice, model construction is a manual process, in which a modeler manually associates a qualitative model with dynamics, and experiments with the resulting model using software tools such as Copasi [4]. Such an approach can give rise to good quality models, but certainly can be seen more as a cottage industry than

P. Lambrix and G. Kemp (Eds.): DILS 2010, LNBI 6254, pp. 164–178, 2010.

as a highly scaleable production process. The widespread use of high throughput experimental methods means that manual modeling can easily become the rate limiting step, and the diversity of available data sets means that modelers operate in a complex information space in which the provenance of a model can be difficult to decipher. As such, there seems to be value in exploring the extent to which the association of models with experimental data – in essence the transition from machines to models to simulations – can be automated.

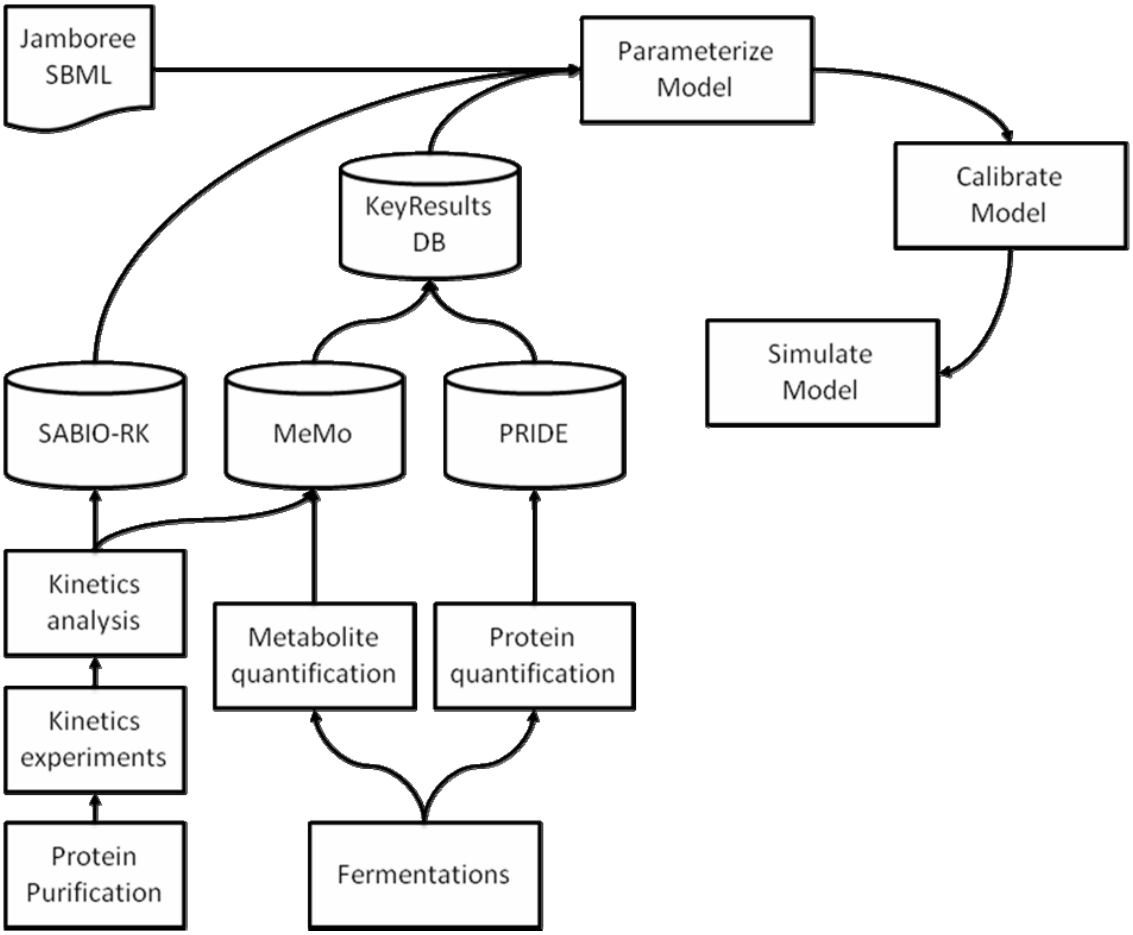

Fig. 1. Overview of key components and the data flows between them

This paper presents an approach to the automation of experimental data capture and integration with models, in which quantitative proteomics, metabolomics and reaction kinetics experimental data is used to parameterize workflows from a metabolic reconstruction, for simulation and analysis using Copasi [4]. In this approach, which is illustrated in Figure 1, the following steps take place:

1. Experimental data is captured directly from instruments, and subject to primary analyses (for example, in proteomics to obtain protein concentrations from mass spectrometry results [5]).

2. Experimental data from instruments, along with the results of the primary analyses, are archived in experimental data repositories, specifically MeMo [6], PRIDE [7] and SABIO-RK [3] that provide a comprehensive record of the experimental processes followed and the results obtained. As such, the experimental data repositories contain the levels of detail that would allow primary analyses such as protein quantifications to be rerun, and the experimental design to be validated.
3. The information required for modeling is extracted from the experimental data resources and stored in a Key Results Database (KRDB), which essentially associates sample information with experimental factors and measured results. Thus the KRDB contains the subset of the data from the experimental repositories at (2) that is required for modeling, and provides consistent representation of quantitative experimental data results for use during model parameterization. As SABIO-RK [3] was essentially already designed to support modeling tasks, we do not replicate reaction kinetics data in the KRDB.
4. A Taverna [8] workflow obtains qualitative model information, represented using SBML [9], parameterizes this model with results in the KRDB, and conveys the resulting quantitative model to the Copasi Web Service [10] for calibration and simulation. An SBML document is built up incrementally through the workflow cycle. Initially an unparameterized, stoichiometric network is extracted from the metabolic reconstruction. The parameterization workflow queries the KRDB to extract initial concentrations of metabolites and enzymes. SABIO-RK is queried allowing each reaction to be expanded in terms of its kinetic equation and kinetic parameters. These parameters are then tuned by the calibration workflow producing a model that may be submitted to the simulation workflow, generating results in SBRML format [11].

The remainder of the paper is structured as follows. Section 2 drills down on the individual components within the lifecycle, describing the key design decisions and the resulting capabilities. Section 3 presents some conclusions on the results to date and their significance for systems biology in practice.

2 Components and Characteristics

2.1 From Equipment to Experimental Results

Three experimental techniques are required to provide data for the parameterization of kinetic metabolic models: quantitative proteomics and metabolomics, and enzyme kinetic assays. In each case there is a requirement to:

(i) perform analyses on the raw experimental data to derive the secondary quantitative parameters required in the model;

(ii) store the raw experimental data along with relevant metadata and the derived parameters, thus providing the facility to trace back and reanalyze raw data should this be required from model simulation results. Where possible, existing data standards and tools are reused, both to reduce wheel-reinvention and development time, and also to provide the facility of sharing experimental data in formats with which the bioinformatics community is familiar.

Quantitative proteomics studies are performed using tandem mass spectrometry, utilizing the QconCAT approach in which isotopically-labelled peptides of known concentration are spiked into a sample, and peak area / intensity comparisons are used to infer peptide and therefore protein concentration in the sample [5].

In order to facilitate the analysis task, a wizard has been produced [12] that automates the steps of:

(i) performing a database search against the Mascot search engine [13] to identify both isotopically-labelled and native peptides;
(ii) determining which peptides can be reliably quantified, based on Mascot significance scores and peptide ion retention times;
(iii) performing the quantification, using an algorithm developed for the SILACAnalyser tool [14];
(iv) formatting both identification and quantification results according to the PRIDE XML data format; and
(v) uploading both the derived results and experimental data to a native XML database.

Following submission to the XML database, data can be extracted and queried via a web and web service interface. It is standard practice in quantitative proteomics for multiple experimental replicates to be performed, and thus quantifications at the protein level are calculated using contributions from individual peptides from each replicate. These can then be queried to parameterize SBML models or extracted to populate the KRDB (see Figure 2).

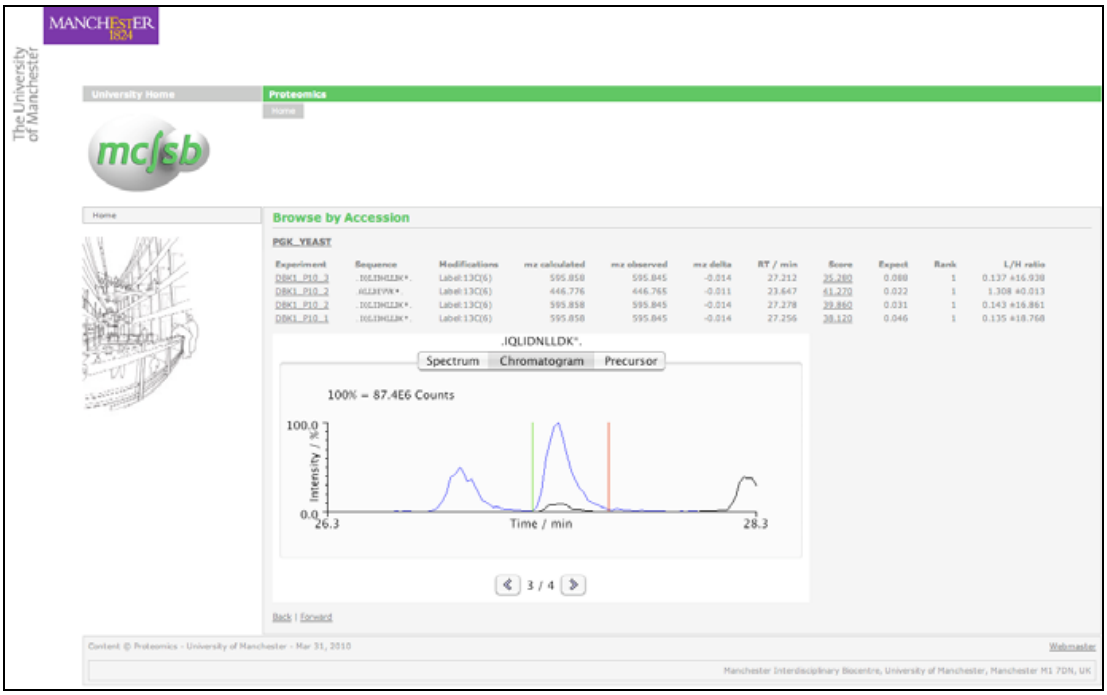

Fig. 2. Web interface displaying raw experimental proteomics data. Absolute protein concentrations can be exported from this data and stored in the Key Results Database.

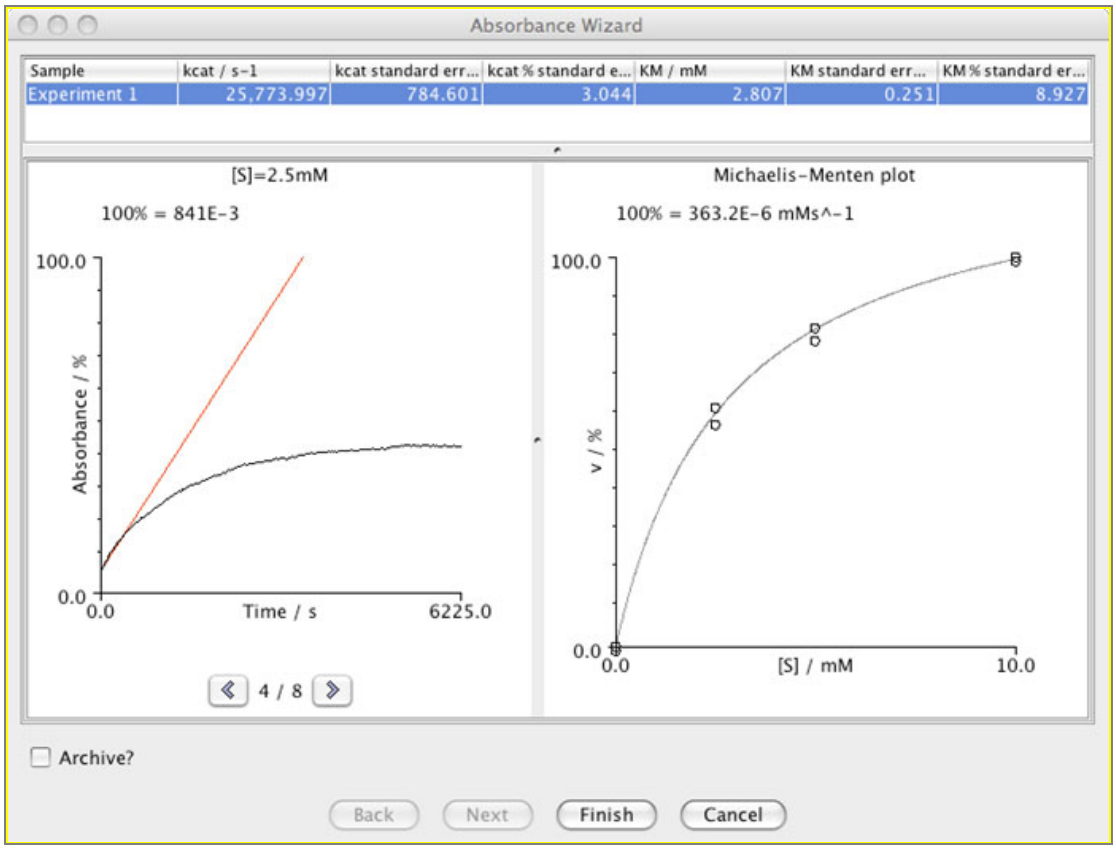

Fig. 3. The KineticsWizard result panel. The left panel displays the raw absorbance data as acquired, upon which the reaction initial rate has been fit. The initial rates from all acquisitions are plotted against substrate concentration in a Michaelis-Menten [17] plot on the right hand side. From this Michaelis-Menten plot, the kinetic parameters k_{cat} and K_M are calculated and displayed in the top panel.

Quantitative metabolomics studies are also performed using tandem mass spectrometry. However, due to the physiochemical diversity of metabolites, these experiments are less homogenous than quantitative proteomics studies, and a range of experimental techniques are performed in order to determine their *in vivo* concentrations. As such, the experimental data analysis step is performed manually, generating a list of metabolite concentrations that can be input into the MeMo database. Metabolite concentrations can then be accessed through a Pierre-generated web and web service interface [15].

Enzyme kinetic assays are performed through spectrophotometry, in which each enzyme of interest is expressed and purified, and the rate of its action measured *in vitro* by measuring the production of reaction product over time.

While the SABIO-RK database is a well-established resource for the storage of kinetic parameters derived from such experiments, there is currently no existing resource for the management of the original time course raw data from which these parameters are derived. As such, a solution has been developed in which derived parameters are stored in SABIO-RK, while associated raw data is managed in an extension of the MeMo database. Both resources are linked through web and web

service interfaces, allowing the user of a given parameter to view and extract the original raw data from which the parameter was derived [16].

With a view to automating experimental data capture, analysis and deposition as much as possible, several wizards have been developed. For example, the KineticsWizard is integrated with the instrument software that performs the tasks of:

(i) calculating kinetic parameters by applying a fitting algorithm to the time course data (see Figure 3);
(ii) capturing sufficient metadata to allow mapping of parameters to models; and
(iii) submitting data to the MeMo and SABIO-RK databases. SABIO-RK provides the facility to export kinetic parameters in SBML format, and can act as a single unified interface both to newly measured in-house parameters and to existing third party kinetic data.

Taken as a whole, the above resources provide the facility for analyzing and managing experimental data in such a way that derived values and parameters may be readily imported into systems biology models.

2.2 From Experimental Results to Key Results

MeMo and PRIDE store experimental data in quite specific formats that preserve information about their acquisition and subsequent processing. Within our workflow, we only require the cellular concentrations of metabolites and enzymes alongside measured kinetic parameters to be able to parameterize our models. The diverse nature of the representation of data in the independently developed experimental results repositories (MeMo, PRIDE, SABIO-RK) lends itself to being consolidated into a single repository to ease interactions with the workflows that parameterize and calibrate the model.

The KRDB [18], the data model for which is illustrated in Figure 4, allows for the amount of metadata associated with a recorded result to be reduced to the minimum necessary to support model development, and thereby facilitates the storage of these results in a set format, no matter which type of experiment they were acquired from. We note that there is not really, therefore, a single *minimum information* requirement for a type of experimental data; rather different users of experimental data results have different requirements. In this context, the target users of the experimental repositories are principally experimentalists who need to understand in detail the process through which results were produced, for example to inform reanalysis. By contrast the target users of the KRDB are modelers, who rarely have the inclination (or perhaps expertise) to make full use of the details captured in the experimental data repositories.

We view a result as a particular reading, or calculated quantity (`Measurement`) of a particular thing (`MeasuredItem, MeasuredItemType`), gathered under a particular set of conditions and possibly at a particular time (in the data model these conditions are referred to as `Factors`). These conditions may be either static

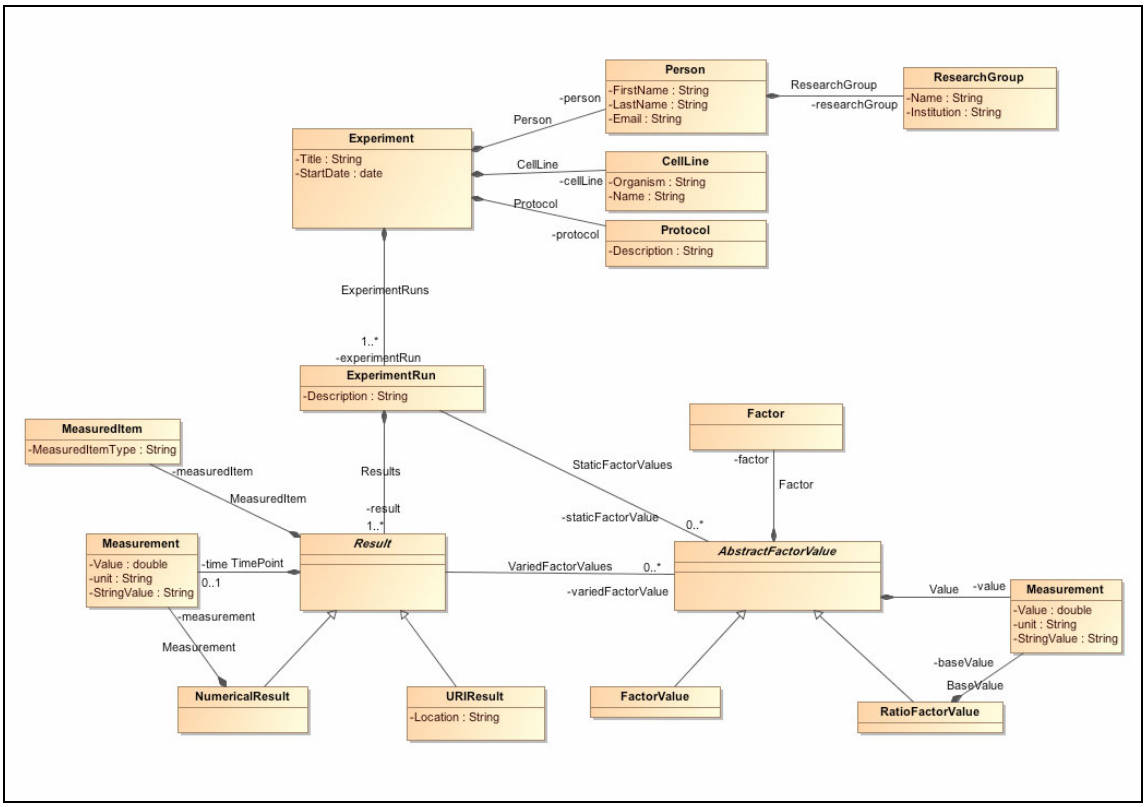

Fig. 4. UML Class Diagram of the Key Results Data Model. Boxes are classes. Arrows indicate classes that have inherited from a parent class. Lines with diamonds indicate classes that are possessed by another class – the diamond end of the line extends from the containing class. Labels indicate what classes represent in the containing class and the cardinality of the relationships. All relationships are 1:1 unless otherwise indicated (0-* is any number of instances or none at all, 1-* is at least one instance).

throughout the experiment (`StaticFactors`) or vary as the readings taken from the experiment progress (`VariedFactors`).

The data model has been converted into an XML schema, and a repository for documents conforming to this schema has been developed. Data for submission to the KRDB is formatted using spreadsheet software into tab delimited files, consisting of a list of measurements and the variable factors associated with each measurement. Once in this format a web form is used to add additional annotation to the experiment and submit the data to the database.

The web form allows basic details of the experiment to be entered, along with any specific conditions surrounding the experiment (the `StaticFactors` described earlier). Upon submission, the form and tab delimited file are checked for basic consistency and then processed by server side software into an XML document conforming to the KRDB schema. This document is then stored in eXist, an open-source, freely available XML database that provides support for web and web service interface development (http://exist-db.org/).

Data may be retrieved from the repository either manually through a web-based front end, or as in the case of our workflows, by using eXist's RESTful web service interface.

For our workflow we store the consolidated cellular concentrations of metabolites and proteins in the KRDB. These numbers are calculated by the experimentalists using the Wizards described in Section 2.1 from their replicate data and recorded in tab delimited format as described above. Metabolites are identified by ChEBI IDs [19], unambiguously linking the metabolite to a defined chemical structure, and likewise enzymes by gene or protein identifiers, such as Saccharomyces Genome Database (SGD) [20] and UniProt [21] respectively. Figure 5 shows a fragment of KRDB XML describing the concentration of the enzyme YCR012W.

```
<CellLine>
    <Name>Y23925</Name>
    <Organism name="Saccharomyces cerevisiae"/>
</CellLine>
<ExperimentRun>
    <TypeOfResult name="Protein Quantification"/>
    <Description>Proteomics quantifications</Description>
    <NumericalResult>
        <VariedFactorValues/>
        <MeasuredItem
             itemType="YeastGeneAccession">YCR012W</MeasuredItem>
        <Measurement unit="cell^-1">2818332.709</Measurement>
    </NumericalResult>
```

Fig. 5. Fragment of KRDB XML showing markup for the quantification of the enzyme YCR012W, measured in the unit copy numbers per cell.

2.3 From Results to Parameterized Models

The three sets of results produced from experiments measuring the activity of enzymes and their concentrations, as well as those of the metabolites involved in enzymatic reactions, are used for the parameterization of systems biology models. Integration of the experimental data with initially qualitative models can be achieved in a systematic manner using procedures constructed and enacted by a workflow management system such as Taverna [8]. These workflows define the flow of data between computational resources, which have been deployed as web services, enabling databases such as SABIO-RK and the KRDB to be queried by the workflow enactment engine.

Taverna workflows have been written to assemble, optimize and simulate parameterized systems biology models. These steps are characterized by successive transformations of a SBML model with quantitative data. Parameterization of a systems biology model initially requires a skeleton SBML model that describes, in a qualitative fashion, the components and their relationships with one another in a biological system. In terms of a metabolic pathway, metabolites and enzymes represent the nodes of this system, whilst the edges between these components represent biochemical reactions. Information about individual metabolic reactions originates from a web service providing access to a metabolic reconstruction. Smaller, more manageable, models comprising specific metabolic pathways are constructed by a qualitative model construction workflow based on some given criteria such as a list of enzyme names, as illustrated in Figure 6. Various metadata are retrieved for enzymes and

metabolites from the consensus metabolic model web service. These metadata include references to external databases including ChEBI and UniProt) identifiers, so that metabolites and enzymes can be uniquely identified within a SBML model. Information representing the association of metabolites and enzymes for each reaction is then retrieved from the metabolic reconstruction web service. Based on this collated data, an SBML document is assembled using classes and methods from libSBML by the qualitative model construction workflow [22].

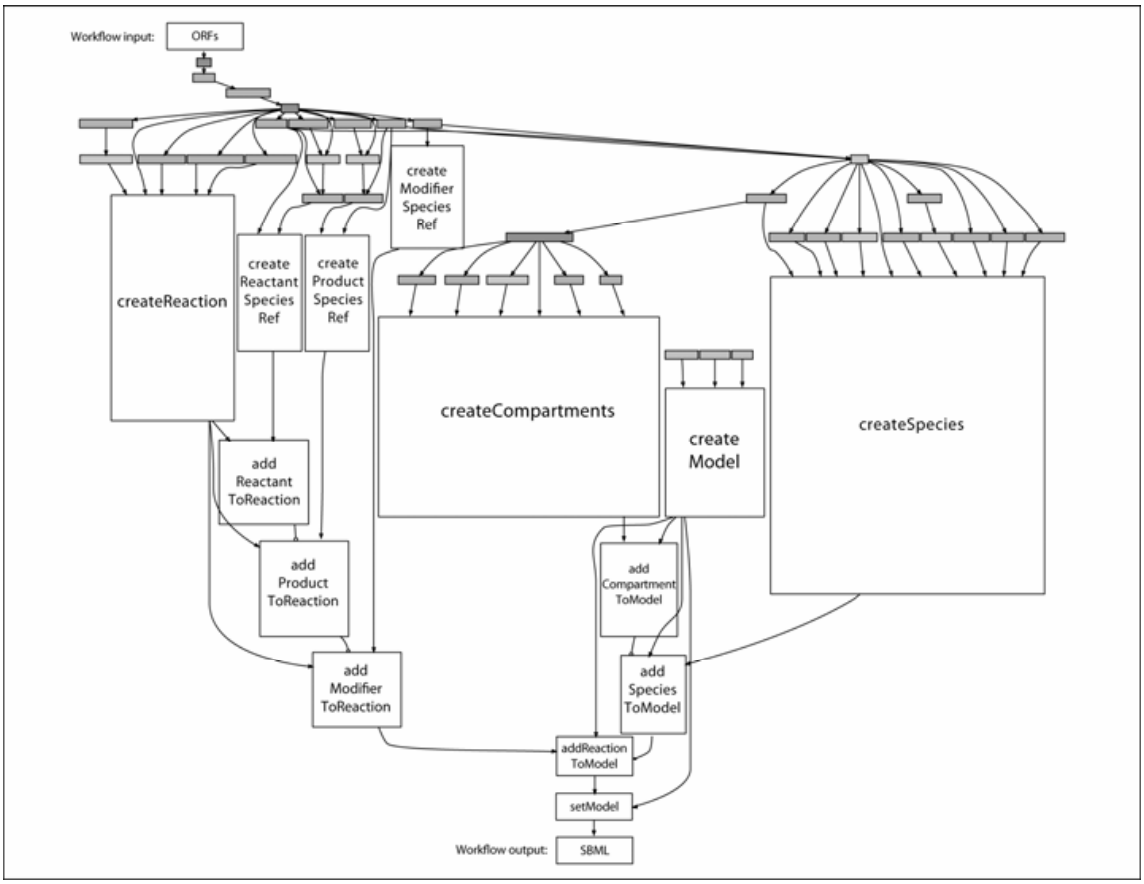

Fig. 6. The workflow used for constructing qualitative models of metabolic pathways in SBML. Calls to the consensus network web service (grey boxes) provide information about the protein, the catalysed reaction and its constituent metabolites for each enzyme from a list of open reading frame numbers. This information is used within nested workflows (white boxes) to iteratively generate components in SBML models using methods from libSBML. An SBML model is produced as the output of the workflow.

The creation of a qualitative SBML model defining how components are related to one another in metabolic reactions provides a context for the integration of proteomics, metabolomics and reaction kinetics data. The parameterization of the SBML model undertaken by this second workflow involves the mapping of quantitative experimental data onto the model which is dependent upon the external database identifiers that have been used to reference metabolites and enzymes in the qualitative SBML model, and the key results and SABIO-RK databases (Figure 7).

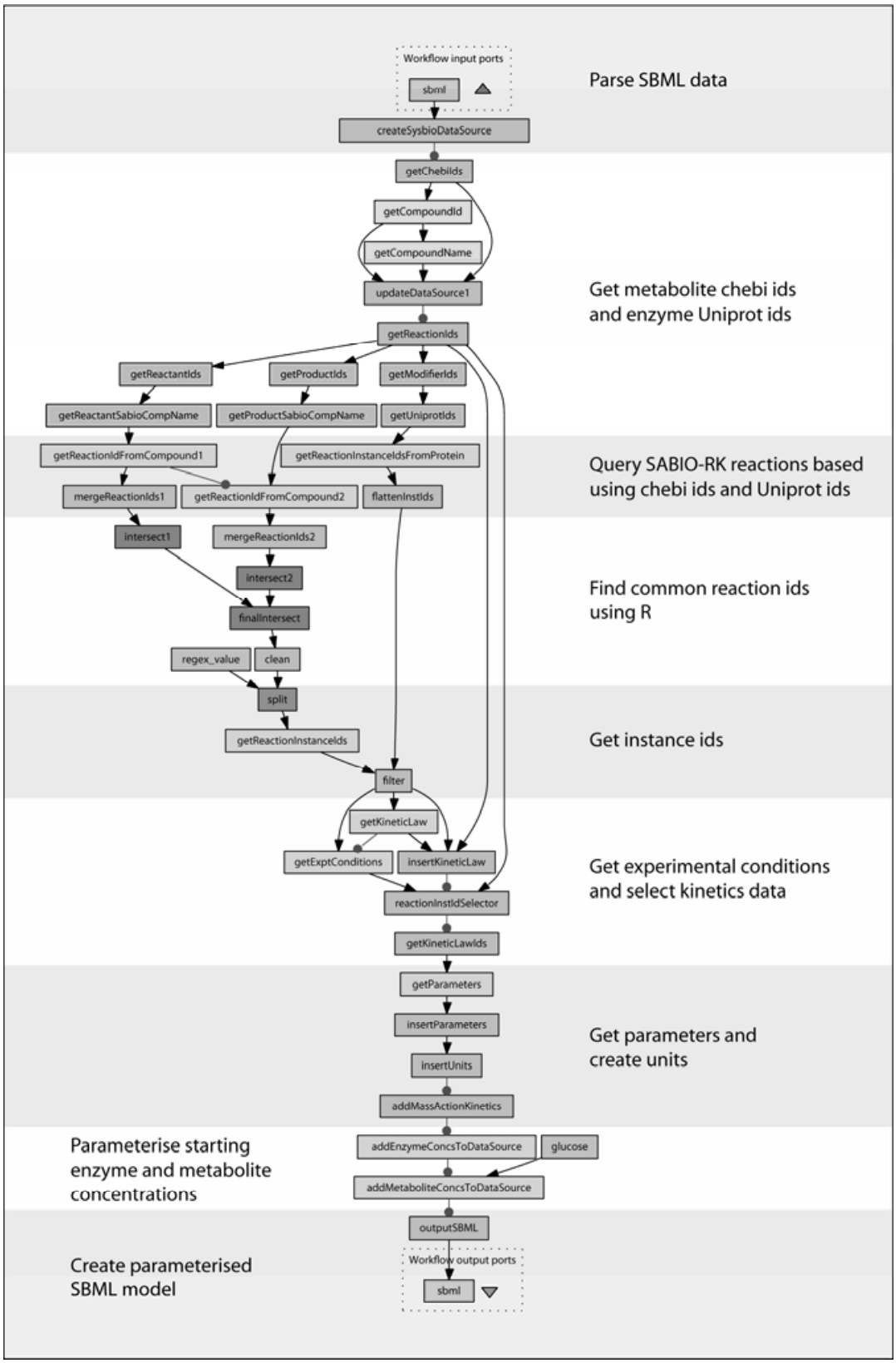

Fig. 7. Model parameterisation workflow integrating experimental data from SABIO-RK and the KRDB with a qualitative SBML model. Quantitative data from SABIO-RK and the KRDB were used to parameterise source metabolites and enzymes with their starting concentrations, and reactions with enzyme kinetics.

```
<sbml>
 <model>
   ...
   <listOfCompartments>...</listOfCompartments>
   <listOfSpecies>
    <species id="M_172" name="ATP" initialConcentration="3.5"/>
    <species id="E_670" name="YCR012W" initialConcentration="0.055110250684886"/>
    ...
   </listOfSpecies>
   <listOfReactions>
    <reaction id="R_1023" name="phosphoglycerate kinase">
     <listOfReactants>
       <speciesReference species="M_4"/>
       <speciesReference species="M_135"/>
     </listOfReactants>
     <listOfProducts>
       <speciesReference species="M_57"/>
       <speciesReference species="M_172"/>
     </listOfProducts>
     <listOfModifiers>
       <modifierSpeciesReference species="E_670"/>
     </listOfModifiers>
    <kineticLaw>
     <math xmlns="http://www.w3.org/1998/Math/MathML">
      <apply>
        <divide/>
          <apply>
            <times/>
              <ci> E_670 </ci>
              <ci> kcat </ci>
              <ci> M_57 </ci>
          </apply>
          <apply>
            <plus/>
              <ci> Km </ci>
              <ci> M_57 </ci>
          </apply>
        </apply>
     </math>
     <listOfParameters>
      <parameter id="kcat" value="343.5"/>
      <parameter id="Km" value="0.77"/>
     </listOfParameters>
    </kineticLaw>
    </reaction>
   ...
</sbml>
```

Fig. 8. A fragment of an SBML model showing the parameterized starting concentration of the enzyme labeled as YCR012W, This concentration was calculated by the parameterisation workflow using the data shown in Figure 5

The starting concentrations of metabolites and enzymes are parameterized with measurements stored in the KRDB by matching ChEBI and UniProt identifiers between data values in this repository with appropriate components in the SBML model (Figure 8). Other sources of data for parameterising starting concentrations of metabolites and enzymes can be used, providing that ChEBI and UniProt database identifiers have been used to reference measurements. In contrast, the combination of ChEBI and UniProt identifiers that defines each metabolic reaction in the qualitative SBML model is used by the parameterisation workflow to search for relevant kinetics in the SABIO-RK database. It is often the case that this search results in multiple sets

of kinetic data being found for a given reaction due to readings that have been measured for enzymes under different assay conditions. In these cases, the parameterisation workflow invites the user to select those kinetics required for the reaction based on experiment conditions. The output of the parameterisation workflow is a SBML document whose reactions have been parameterised with reaction kinetics and starting concentrations for source metabolites.

2.4 From Parameterised Models to Simulation Results

Prior to their use in predictive studies, parameterized models may be optimized in order to improve their accuracy when used in simulations. Experimental measurements of metabolite concentrations can be used to modify parameters in reaction kinetics until the output of the model produces results similar to those obtained from experimentation. Optimization of the SBML model can be performed in workflows by making use of an optimization algorithm in COPASI that has been exposed as a web service [10]. However, optimization of an SBML model is a complicated process. Firstly, there is the problem with mapping metabolomics measurements with metabolites in systems biology models. Secondly, the process of model optimization requires selection of those parameters to optimize and to what extent. A model calibration workflow has been implemented which converts experimental data into SBRML, thus allowing metabolomic measurements to be associated with components in SBML models [11]. This workflow also features the use of a pop up window wizard that invites the user to configure those parameters requiring optimization.

The calibration workflow uses the COPASIWS optimization web service that has been implemented in an asynchronous fashion due to the compute-intensive nature of the process. The workflow initiates a request for a job identifier that is then used to ensure that data is loaded and configured appropriately for each optimization process. The output of the calibration workflow is a SBML model whose original reaction parameter values have been modified against metabolomics measurements by the COPASIWS optimization web service. This optimized SBML model may now be used for simulation, which can also be performed using workflows. Such a workflow was constructed which invokes the time course simulation service available from COPASIWS, which can provide results in SBRML format for further processing (Figure 9). For example, the time course results for specific metabolites can be extracted and then used to plot how concentration varies according to time.

Whilst workflows can automate the integration of data, manual perusal of the models between each workflow stage of the process of generating optimized systems biology models is required to ensure that they make sense from a biological point of view. The set of systems biology workflows is reliant on data and metadata in all databases being consistent, otherwise anomalous models can be generated. For example, the presence of charge-balancing protons in reactions from one database but not for the same reaction in another source can lead to the inability of workflows to parameterise reactions. Problems with parameterisation of starting concentrations can also occur when the same metabolites have been referenced with different ChEBI identifiers in databases.

```
<?xml version="1.0" encoding="UTF-8"?>
<sbrml version="1" level="1" creationDate="2010-03-31">
  <model name="Test model" sourceURI="..."/>
    <operations>
      <operation id="op1" name="Time Course" ontologyTerm="term2">
        <method name="Deterministic (LSODA)" ontologyTerm="term1"/>
          <software name=" COPASI" version="COPASI 4.5 Build 30" />
            <result>
              <resultComponent id="component1">
                <dimensionDescription>
                  <compositeDescription name="Time" ontologyTerm="term3" indexType="double">
                    <compositeDescription name="Species" indexType="string">
                      <tupleDescription>
                        <atomicDescription name="Concentration" ontologyTerm="term4"
                               valueType="double"/>
                        <atomicDescription name="Particle Numbers" ontologyTerm="term5"
                               valueType="Integer"/>
                      </tupleDescription>
                    </compositeDescription>
                  </compositeDescription>
                </dimensionDescription>
                <dimension>
                  ...
                  <compositeValue indexValue="40">
                    <compositeValue indexValue="M_273">
                      <tuple>
                        <atomicValue>0</atomicValue>
                        <atomicValue>0</atomicValue>
                      </tuple>
                    </compositeValue>
                    <compositeValue indexValue="M_292">
                      <tuple>
                        <atomicValue>3.5</atomicValue>
                        <atomicValue>2.10775e+24</atomicValue>
                      </tuple>
                    </compositeValue>
                    ...
</sbrml>
```

Fig. 9. A fragment of an SBRML file showing the results in the form of changes in metabolite concentrations produced by the time course simulation workflow

3 Conclusions

The construction of predictive metabolic models from experimental data is a considerable bottleneck in high throughput Systems Biology. Our information management strategy has defined a process (Figure 1) whereby data may be acquired and used in a systematic, and largely automated way, go some way towards alleviating this limiting step.

The initial hurdle for any information management strategy is not the development of repositories, but the streamlining of the process of data acquisition in such a way that those generating it do not perceive it as burdensome. Adding value during the process of data acquisition shifts the benefits from some detached individual somewhere down the line to one that is of obvious relevance to the experimentalist. Our "wizard" acquisition tools perform precisely this function; the PrideWizard, as well as capturing metadata associated with experiments, automates the quantification of peptides within minutes rather than the days it takes to perform this task by hand. The KineticsWizard also captures metadata, but automates the calculation of kinetic parameters, removing the need to perform this task by hand.

Once acquired, the data must be made available to consumers. These consumers will either be modelers or, in the case illustrated in this paper, modeling workflows.

We are aware that the consumers may not have access to the resources needed to implement or maintain the heavyweight experimental databases (MeMo, PRIDE) that we have used to archive our data. To this end, the KRDB was developed to integrate data from MeMo and PRIDE, and make it available to systems biology workflows. The use of the KRDB makes our data and workflows both useful to and implementable by others. Utilisation of such generic resources ensures that the system is applicable to a range of organisms. Although yeast is studied in this demonstration, none of the tools used are limited to this organism.

The Taverna workflows that assemble, optimize and simulate systems biology models are the consumers of the acquired data, and are the culmination of the information management workflow. By automating these processes we have tackled the perceived rate-limiting step of model construction, and provided the facility to run repeated simulations incorporating automatically selected parameter values.

The facility to simulate multiple parameter sets quickly is valuable. It allows multiple hypotheses to be tested *in silico,* which may then inform experiments that need to be conducted in order to validate the generated models, or highlight elements of metabolism that may be manipulated experimentally to further understanding.

As high-throughput experimental techniques continue to improve, the problem of how to manage the data generated will be a perpetual one. The workflow, tools and repositories presented and described here demonstrate how integrated information management can support and expedite the processes of integrative systems biology, something that can only become of increasing importance.

Availability

All workflows and accompanying documentation are available from myExperiment at http://www.myexperiment.org/packs/107. The Taverna workbench (version 2.1) can be downloaded from http://www.taverna.org.uk to run workflows, which make use of a key results database available from http://beaconw.cs.manchester.ac.uk:8780/mcisbkrdb/ and SABIORK that is accessible at http://sabio.villa-bosch.de. The COPASI web service is available from http://www.comp-sysbio.org/CopasiWS/.

Acknowledgements

The authors thank the EPSRC and BBSRC for their funding of all authors through the Manchester Centre for Integrative Systems Biology grant (BBSRC/EPSRC Grant BB/C008219/1) and thank the BBSRC for the funding of Daniel Jameson through the Dynamics and function of the NF-κB signalling system SABR grant (BB/F005938, BB/F00561X).

References

1. Herrgard, M.J., et al.: A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat. Biotechnol. 26(10), 1155–1160 (2008)
2. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems Biology in practice. Wiley, Chichester (2005)

3. Rojas, I., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Wittig, U.: Storing and annotating of kinetic data. Silico Biol. 7 (suppl. 2), S37–S44 (2007)
4. Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J., Kummer, U.: Computational modeling of biochemical networks using COPASI. Methods Mol. Biol. 500, 17–59 (2009)
5. Rivers, J., Simpson, D.M., Robertson, D.H., Gaskell, S.J., Beynon, R.J.: Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. Mol. Cell. Proteomics 6(8), 1416–1427 (2007) (Epub May 17, 2007)
6. Spasic, I., et al.: MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. BMC Bioinformatics 7, 281 (2006)
7. Jones, P., Côté, R.G., Martens, L., Quinn, A.F., Taylor, C.F., Derache, W., Hermjakob, H., Apweiler, R.: PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res. 34, D659–D663 (2006)
8. Oinn, T., et al.: Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20, 3045–3054 (2004)
9. Hucka, M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4), 524–531 (2003)
10. Dada, J.O., Mendes, P.: Design and Architecture of Web Services for Simulation of Biochemical Systems. In: Paton, N.W., Missier, P., Hedeler, C. (eds.) DILS 2009. LNCS, vol. 5647, pp. 182–195. Springer, Heidelberg (2009)
11. Dada, J.O., Spasic, I., Paton, N.W., Mendes, P.: SBRML: a markup language for associating systems biology data with models. Bioinformatics (Feburary 21, 2010) (Epub ahead of print)
12. Siepen, J.A., et al.: An informatic pipeline for the data capture and submission of quantitative proteomic data using iTRAQTM. Proteome Sci. 5, 4 (2007)
13. Perkins, D.N., et al.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20(18), 3551–3567 (1999)
14. Nilse, L., et al.: SILACAnalyzer - a tool for differential quantitation of stable isotope derived data. In: CIBB, 6th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Genoa (2009)
15. Garwood, K., et al.: Model-driven user interfaces for bioinformatics data resources: regenerating the wheel as an alternative to reinventing it. BMC Bioinformatics 7, 532 (2006)
16. Swainston, N., et al.: Enzyme kinetics informatics: from instrument to browser. FEBS J. (submitted 2010)
17. Michaelis, L., Menten, M.L.: Die Kinetik der Invertinwirkung. Biochem. Z, 49, 333–369 (1913)
18. Jameson, D., et al.: Lightweight Experimental Data Management for Systems Biology (submitted 2010)
19. Degtyarenko, K., et al.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–D350 (2008)
20. Issel-Tarver, L., et al.: Saccharomyces Genome Database. Methods Enzymol. 350, 329–346 (2002)
21. Schneider, M., et al.: The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. J. Proteomics 72, 567–573 (2009)
22. Li, P., et al.: Automated manipulation of systems biology models using libSBML within Taverna workflows. Bioinformatics 24, 287–289 (2008)