Special issue papers

Simultaneous modelling of metabolic, genetic and product-interaction networks

N. S. Juty, H. D. Spence, H.-R. Hotz, H. Tang, I. Goryanin and T. C. Hodgman

Date received (in revised form): 8th June 2001

Navtej S. Juty, Hugh D. Spence, Igor Goryanin and **Charlie Hodgman**
are in the Medicines Research Centre at GlaxoSmithKline Pharmaceuticals. Their research interests include process informatics, cellular simulations and the modelling and analysis of complex systems.

Hans-Rudolf Hotz
is a database programmer at the Sanger Centre, Cambridge.

Haizhou Tang
is a postgraduate student at the University of Manchester whose research interests include the integration of transcriptome, proteome and metabolome data for the simulation of physiological processes in model organisms.

Keywords: *whole cell model, pathway, metabolism, protein-interaction networks, genetic network, XML*

Abstract

The creation of cell models from annotated genome information, as well as additional data from other databases, requires both a format and medium for its distribution. Standards are described for the representation of the data in the form of Document Type Definitions (DTDs) for XML files. Separate DTDs are detailed for genetic, metabolic and gene product-interaction networks, which can be used to hold information on individual subsystems, or which may be combined to create a whole cell DTD. In the execution of this work, a fifth DTD was also created for a metabolite thesaurus, which allows incorporation of metabolite synonyms and generic nomenclature data into the models. A gene-regulation classification scheme was also created, to facilitate incorporation of gene regulatory information in an efficient manner. The work is described with particular reference to the metabolic network of *Escherichia coli*, which contains 808 individual enzymes. The assignment of confidence levels to these data, through the use of Gene Ontology evidence codes, is highlighted. *In silico* investigations may now be performed using the mathematical simulation workbench, DBsolve, which incorporates the facility to introduce data directly from XML.

INTRODUCTION

Since the sequencing of *Haemophilus influenzae*,[1] the first bacterial genome, an additional 50 have been added to those publicly available. A further 15 are currently being annotated, while another 86 are being sequenced.[2] Clearly, the phenotype of an organism is a result of its genotype, anatomy and the surrounding environment. Thus, if the genome sequence of an organism has been determined, an *in silico* model may be reconstructed to assemble the components into a functional, virtual organism. Given also a defined set of environmental conditions, one can begin to make predictions regarding cellular activities in this context.

This constructionist, integrative approach would translate biochemical knowledge into mathematical formulae, thereby permitting a meaningful interpretation of genome-scale information. This represents the next logical step for data analysis in the 'post-genome' era.

Current research directions

Various groups are attempting reconstruction of some aspect of cellular function, though most activity seems concentrated on metabolic networks and their mathematical modelling.[3–6] Metabolic modelling and pathway analysis are intensive areas of research, yielding new analytical theories and methodologies.[7] Recently, an *in silico* reconstruction of the *Escherichia coli* metabolic network gave results consistent with those observed experimentally, *in vitro*.[8] Other groups are taking a more

N. S. Juty,
GlaxoSmithKline Pharmaceuticals,
Medicines Research Centre,
Gunnels Wood Road,
Stevenage SG1 2NY, UK

Tel: +44 (0) 1438 763878
E-mail: ncj10219@gsk.com

holistic view, attempting to model more than one subsystem. Of these, E-cell[9,10] currently appears to lack sufficient implementation of mathematical algorithms to make thorough analysis feasible, while V-cell[11] maps only specific cellular processes onto electronic subcellular images.

Data validation is a crucial step

The choice of *E. coli*

Despite the relatively large size of the genome compared to other bacterial species, such as *Mycoplasma genitalium*,[12] *E. coli* is a suitable first organism with which to attempt the creation of a virtual cell. Besides the availability of the genome sequence,[13] there is a wealth of additional information about this prototroph that is freely accessible. While SWISS-PROT[14] provides a good general source of information on the proteins found in *E. coli*, metabolic and pathway reconstruction data are available at WIT,[15] KEGG[16] and EMP.[17] In addition, the ENZYME[18] and BRENDA[19] databases provide specific enzyme-related reports. Information pertaining to *E. coli* genes is also available at GenProtEC,[20] while RegulonDB[21] presents information about the regulation of these genes and operons. As more data are evaluated, these too are being made available on the Internet.[22]

***E. coli* data are freely available and easily accessible**

***E. coli* is widely used in industry**

Besides its status of being the best-studied microorganism, and the abundance of published data available, *E. coli* is also widely used in the pharmaceutical industry to generate both drug precursors[23] and vaccines[24,25] through various fermentations[26,27] and also has other widespread commercial application.

For these reasons, our initial attempts to create cellular models are targeted to prokaryotic organisms, beginning with *E. coli*, although the data formats described below have been created with both prokaryotes and eukaryotes in mind.

INFORMATION FROM THE WEB AND DATA HANDLING

The importance of validation

Unfortunately, the abundance of data related to *E. coli* also has the added drawback that erroneous data entries tend to be propagated in the various databanks.[28] This makes data validation a crucial step in creating an accurate model.

Biochemical verification lags a long way behind inferred function, but models relying solely on the highest quality data would be too incomplete to be of use. Therefore models do require the inclusion of uncertain data, but some assessment of the confidence of every activity and association must be made. We have adopted the evidence codes used by the Gene Ontology Consortium (GO),[29,30] which range from 'Traceable Author Statement' (TAS – where the activity has been confirmed and published) through decreasing levels of inference to 'Non-traceable Author Statement' (NAS).

This data validation, though manually intensive, provides three major functions:

- A mechanism to quantify uncertainty in the model.
- A means to identify the enzymes, etc., that require further biochemical characterisation.
- An approach to enable the computer to selectively delete the more uncertain components of the model.

Structuring available knowledge

Representing the knowledge required to encapsulate even the simplest biological process is challenging, but achievable. Ideally the scientific community would use a single scheme, thereby allowing exchange of model data between various groups. To facilitate this dissemination of information, a suitable medium such as eXtensible Markup Language[31] (XML) is required. XML is already used by the mathematics[32] and chemistry[33] communities, as well as having been adopted by the Gene Ontology Consortium (for a recent review, see Achard *et al.*[34]). The structure of an XML document is defined by its Document

Type Definition (DTD), or alternatively by an XML schema. By comparison to the actual XML file, the DTD or schema itself is relatively small and easily distributed. Although an XML schema is undoubtedly superior for structure and content definition, particularly with regard to the provision of richer datatyping, it is larger and less immediately interpretable to the reader. In addition, the official XML schema specification has only recently been finalised.[35] It may be for these reasons that DTDs are still more commonly used than schemas in defining XML files and their content. In any case, DTDs and schema are interconvertible (though with some limitations) and we envisage provision of our DTDs as schema at a later date.

eXtensible Markup Language is a common medium for sharing information

We have developed DTDs suitable for representing metabolic, genetic and product-interaction networks, whole cell model data, as well as a metabolite thesaurus. The relationship between the metabolic XML file and its DTD is illustrated in Figure 1, as an example. By making our DTDs freely available (see below) we hope to make significant strides in standardising a suitable format for whole cell modelling. Though the DTDs

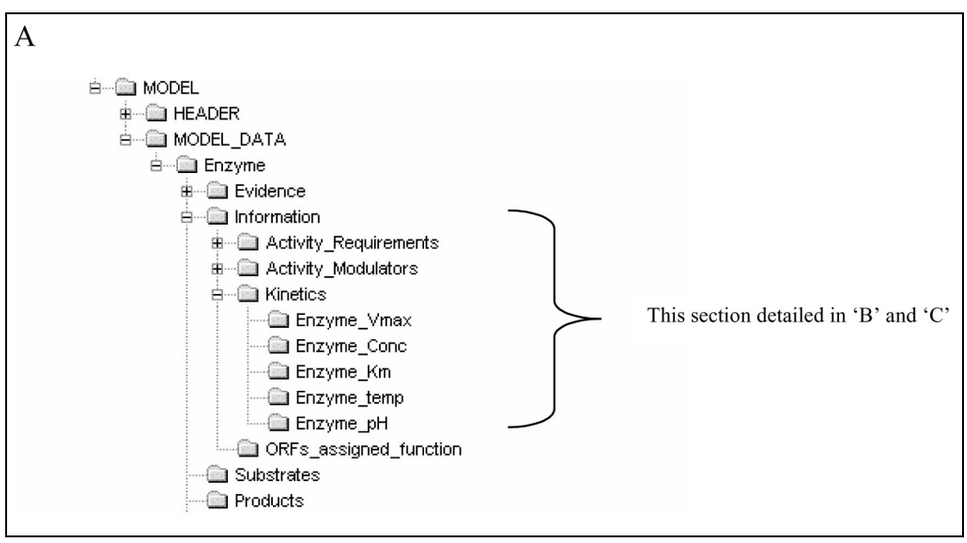

Figure 1: The relationship between DTD and XML. (A) The DTD can be shown as a tree structure, where individual branches represent ELEMENTs in the DTD. The root of the tree ('MODEL') is the parent of 'HEADER' (which contains data not required to generate a simulation) and 'MODEL_DATA'. 'Enzyme', the child of 'MODEL_DATA', has been expanded in part, allowing one to see the parent–child relationships in the subsection indicated (right brace, see 'B'). Branches that may be expanded further are indicated with a '+' left of the ELEMENT folder. (B) The expected frequency with which ELEMENTs will be encountered in a valid XML file is shown. The ELEMENT 'Information' may have the child ELEMENTs 'Kinetics' and 'reaction_reversible', among others (not shown). The optional occurrence of an ELEMENT is given by '?' following the ELEMENT name. In addition, ELEMENTS encountered one or more times ('+'), zero or more times ('*'), or only once (no suffix) can all be described. ELEMENTs may also have associated attributes. The expanded section (see section 'C') illustrates the use of attributes (a list of which falls under '!ATTLIST') to describe the 'substrate' and 'species' from which the V_{max} data are derived. Identically named attributes are also available for use by other kinetic data such as K_m and pH (not shown). (C) A section of XML file described by the DTD is shown. Two periods ('..') have been used to indicate the omission of data (to aid clarity). The nested nature of the XML file mirrors that of the DTD shown above. In this instance, the various 'Kinetics' child ELEMENTs each occur once, though multiple instances are allowed. Using the attributes 'species' and 'substrate' to distinguish between them, this allows the capture of multiple kinetic values from different related (or unrelated) species

DTDs can be visualised as tree structures

B

```
<!ELEMENT Information (.. Kinetics?, reaction_reversible?..)>
<!ELEMENT Kinetics (Enzyme_Vmax*, Enzyme_Conc*, Enzyme_Km*, Enzyme_temp*, Enzyme_pH* ..)>
<!ELEMENT Enzyme_Vmax (Enzyme_Vmax_value?, Enzyme_Vmax_comment?, Enzyme_Vmax_ref?, Enzyme_Vmax_URL?)>
<!ELEMENT Enzyme_Vmax_value (#PCDATA)>
<!ELEMENT Enzyme_Vmax_comment (#PCDATA)>
<!ELEMENT Enzyme_Vmax_ref (#PCDATA)>
<!ELEMENT Enzyme_Vmax_URL (#PCDATA)>
<!ATTLIST Enzyme_Vmax
    species CDATA #IMPLIED
    substrate CDATA #IMPLIED>
```

C

Kinetic data from other organisms can also be captured

```
<Information>
     <Activity_Requirements>..
     </Activity_Requirements>
     <Activity_Modulators>..
     </Activity_Modulators>
     <Kinetics>
          <Enzyme_Vmax species="ovis_aries" substrate="unknown">
               <Enzyme_Vmax_value>0.475</Enzyme_Vmax_value>
               <Enzyme_Vmax_comment>To be filled in by hand</Enzyme_Vmax_comment>
               <Enzyme_Vmax_ref>biochem.,1985,24(19),5099-5106</Enzyme_Vmax_ref>
               <Enzyme_Vmax_URL>www.nottelling.com/=SEL89293-01</Enzyme_Vmax_URL>
          </Enzyme_Vmax>
          <Enzyme_Km species="escherichia_coli" substrate="L-arginine">
               <Enzyme_Km_value>0.012</Enzyme_Km_value>
               <Enzyme_Km_comment>To be filled in by hand</Enzyme_Km_comment>
               <Enzyme_Km_ref>biochem.,1988,27(17),6343-6348</Enzyme_Km_ref>
               <Enzyme_Km_URL> www.nottelling.com/=PRO96327-02</Enzyme_Km_URL>
          </Enzyme_Km>
          <Enzyme_temp species="escherichia_coli">
               <Enzyme_temp_value>37</Enzyme_temp_value>
               <Enzyme_temp_comment>no comment</Enzyme_temp_comment>
               <Enzyme_temp_ref>biochem.,1988,27(17),6343-6348</Enzyme_temp_ref>
               <Enzyme_temp_URL>www.nottelling.com/=PRO96327-02</Enzyme_temp_URL>
          </Enzyme_temp>
          <Enzyme_pH species="escherichia_coli">
               <Enzyme_pH_value>8.1...8.5</Enzyme_pH_value>
               <Enzyme_pH_comment>Optimal pH (range)</Enzyme_pH_comment>
               <Enzyme_pH_ref>biochem.,1979,18(14),3171-3178</Enzyme_pH_ref>
               <Enzyme_pH_URL>www.can'ttellyou.come/=TAT00338-01</Enzyme_pH_URL>
          </Enzyme_pH>
          ..
     </Kinetics>
     <reaction_reversible>NO</reaction_reversible>
</Information>
```

Figure 1: (*Continued*)

themselves are rigid in structure, we envisage periodic updates, and backward compatibility, to accommodate new information in the associated XML files. Example XML files to demonstrate how a document should be filled, as well as the DTDs, are now available (see below).

CREATION OF MODEL FILES

Cell complexity is decomposed into 3 subsystems and subsequently reintegrated

To unravel the complexity of the cell we have decomposed the problem, initially, into three subsystems; metabolic, genetic and (gene) product-interactions, which have historically been modelled in isolation anyway.[36,37] Each system has an associated DTD, all of which can be combined to create a whole cell DTD format. Compliant XML files can therefore be created for each subsystem, or may be combined to generate a whole cell model file.

In each case the information gathered from the various databases is used to create an XML file. This file is then translated into a stoichiometric matrix containing the entity information. For example within the metabolic network

matrix, each enzyme reaction occupies a row, and the columns correspond to compounds or metabolites which are acted upon. The catalytic conversion of one compound (substrate) to another (product) is represented by negative or positive integer values (respectively), where the numerical value itself represents the stoichiometry of the reaction. A reversible reaction would therefore require two rows, one each for the forward and back reactions. In addition, each metabolite in the matrix is assigned a compartment, allowing one to model the action of transporters. These would act by altering the compartment assignment of the metabolite. For example, $glucose_{extracellular}$ would become $glucose_{cytoplasm}$ following the action of a glucose transporter.

Data tracking allows subsequent reassessment

Metabolic network

Initially, it seems prudent to assemble each of the three networks individually. Following validation of the data contained within each network XML file, the individual matrices may be merged. As a first step, we have created a metabolic XML file compliant with the metabolic DTD. Most data were directly extractable from WIT, KEGG and EMP databases, though manual addition was required to fill some information such as GO evidence codes, for each individual enzyme activity. The 'root' of the whole cell XML file consists of a list of the genes. The major ELEMENT of the metabolic network, a subset of the whole cell DTD, is the 'Enzyme', while 'Interactions' is the major ELEMENT of the product-interaction network (represented in Figure 2). Within 'Enzyme', for example, information is held regarding enzyme identifiers (such as Blattner number, gene name and SWISS-PROT identifiers), subunit composition, reaction(s) catalysed and reaction kinetics (Figure 1). Although kinetics and some other fields are currently unused, the format and facility do exist to capture and represent this information. These data can be employed once the static model has been refined to a satisfactory degree. Where information relating to the subject organism is unavailable, alternative values are taken from related species. This situation occurred frequently for specific kinetic information, such as K_m and V_{max}, where values were often available only from non-target (in this case non-*E. coli*) organisms. Importantly, all (kinetic) data entries are associated with attributes describing the organism and substrate for which the entry was recorded, and a reference Element detailing the information source. This allows tracking of data that have been entered in the file, whose accuracy may need to be reassessed at a later time.

Base unit of the metabolic network is 'ENZYME'

Genetic network captures the immediate Protein-DNA interaction involved in the control of gene expression

Various levels of gene expression can be represented

Genetic network

The genetic network represents the regulatory systems involved in the control of gene expression. For transcriptional regulation (the major regulatory mechanism in prokaryotes), the interaction of sigma factor with RNA polymerase is essential. Interactions required for activation or repression are mediated by regulators, which may or may not require one or more cofactors. It is entirely feasible and reasonable to create a genetic regulation classification system based on these observations (see below). Within our scheme, regulations are classified by the number of regulators and cofactors involved, as well as by the effect of the interaction between gene and regulator, and between regulator and cofactor. This simplified overview can be used to depict Boolean regulatory networks where expression is either switched on or off. To create a more representative scheme, we have incorporated the ability of genes to be either basally expressed, or not. A basally expressed gene requires only the presence of the sigma factor in association with RNA polymerase at the promoter, to give a fixed amount of expression. A numerical value can be fixed for this basal expression level by the investigator (presumably based on experimental data). Various degrees of gene expression can

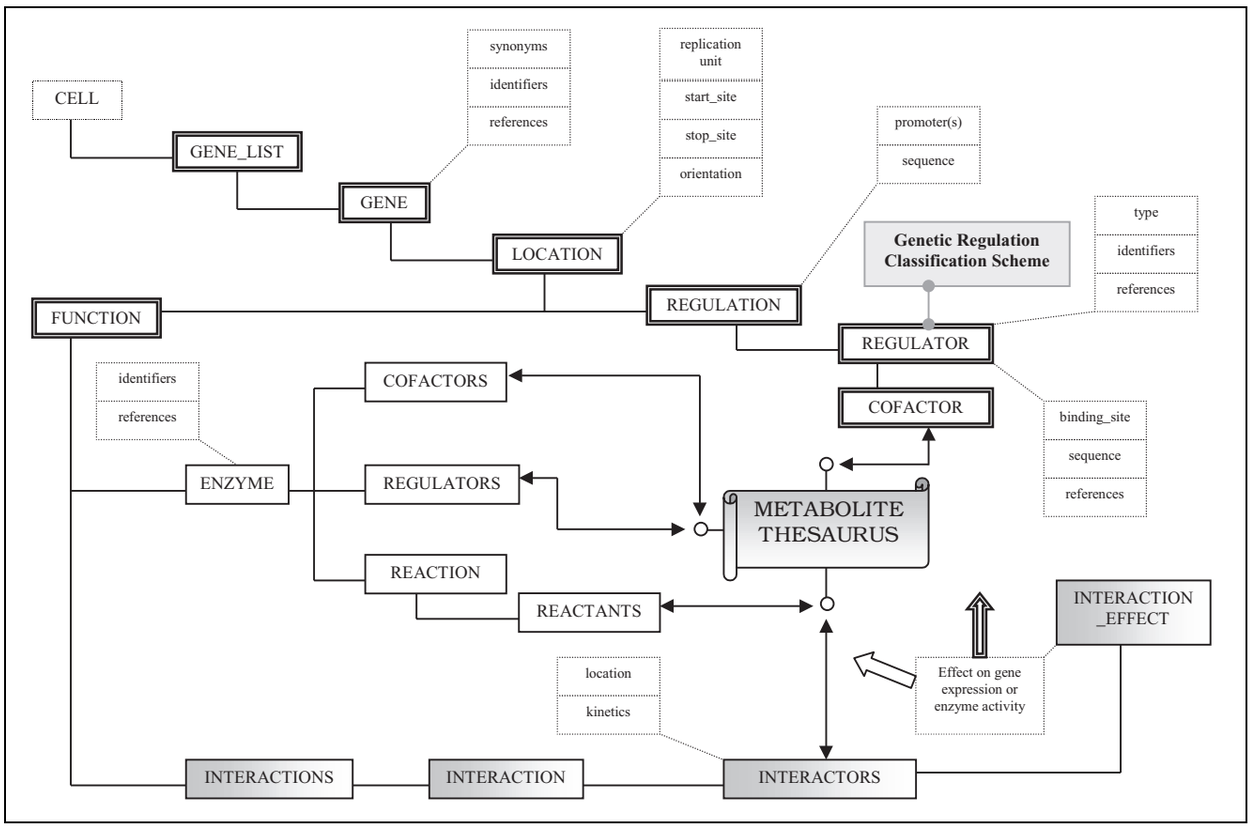

Figure 2: Simplified representation of some DTD components of the integrated virtual 'CELL' model. The metabolic (regular box, no fill), genetic (double-line box, no fill) and product-interaction (regular box, horizontally fading grey fill) DTD ELEMENTs are depicted. Their interrelationships, some of which are mediated through a genetic regulation classification scheme (regular box, grey fill) or a metabolite thesaurus ('scroll'-box, vertically fading grey fill), are also indicated. Some examples of various subELEMENTS are illustrated by unfilled, dashed boxes. Blocked arrows indicate some product-interaction network influences on the metabolic (regular line) and gene regulation (double-line) networks. For a complete description of each component DTD, refer to the complete DTD listings[38]

Product-interaction network captures processes such as protein kinase cascades and multimolecular complexes

also be included as required, simply by adding extra rows in the matrix, where the presence of additional regulators can also be represented.

The creation of this scheme allows representation of an actual regulatory event for a gene simply by provision in the XML file of a reference class name for the regulation, together with a simple list of the regulators and cofactors involved in the regulation. For eukaryotic gene expression, the sigma factor would be replaced with a transcription factor, and additional regulation classes would need to be included.

Product-interaction network

The gene product-interaction network represents the interactions between protein and/or RNA species that transfer information from one part of the cell to another and effect macromolecular synthesis. This would include processes such as protein kinase cascades, chemotaxis, transcription and translation. For each gene product, there may be a number of possible interactions with other entities. Each of these in turn may affect the activity of one or other interactants, or both may associate together to form a new entity. This in turn may be capable of further interactions, or may associate with other entities, for example, to form an assembly or construct (structural or with a defined activity). Entities present in this matrix will often exist already in at least one of the previous networks, either implicitly or by association.

Metabolite thesaurus

Inconsistent nomenclature makes a metabolite thesaurus essential

The most significant hurdle to surmount is the nomenclature used to describe compounds and metabolites. The task of creating a stoichiometric matrix where compounds are not duplicated (or quadruplicated) is not trivial. Inaccuracies may result from synonym use, from typographical error, or may even be an effect of generic-nomenclature usage, for example 'amino acid' for alanine, cysteine, etc. All these factors must be contended with in the creation of an accurate and meaningful model. For example, incorporation in the matrix of an enzyme acting on an alcohol requires the software to recognise the multiple identities associated with such a generic term. To overcome these problems, our methodology incorporates a metabolite thesaurus that will replace synonyms with their accepted 'proper' name, and also replaces generic compounds with a range of accepted compound terms (represented in Figure 2). This is crucial for accurate interconnections to be made between the networks, where a compound such as cAMP (adenosine-3′,5′-cyclic monophosphate or cyclic adenylic acid) may be present in more than one.

SOFTWARE TOOLS TO WORK WITH XML

The cellular models constitute not only the information contained in the XML files, but also the algorithms used to perform the analyses. Our previously described software, DBsolve 5.00,[39] is being modified to read, create and run XML models. DBsolve 6.00 also includes the facility to introduce data directly from XML files and implements algorithms for flux balance analysis (FBA),[40] strain optimisation[41] and elementary modes analysis.[42]

Escherichia coli model

The *E. coli* model has > 3600 reactions

Currently the metabolic model comprises 808 enzymes (3651 reactions in total, involving ~1800 metabolites) for which confidence-codes have been assigned (Table 1). This represents automatically retrieved entries, as well as those orphan[43] activities appended to the file manually. Almost a fifth of entries (17.5 per cent) are fully validated, while over a quarter are confidently assigned ('IDA' and 'IMP'). Approximately a fifth of the enzymes have little or no evidence supporting their biochemical activity ('NAS'). Given that much gene functional assignment is based on sequence similarity, these results are quite encouraging. The metabolic model file, including appendixed orphan activities, is approximately 4 Mb in size. This size is expected to increase as more activities are added, and following the action of the metabolite thesaurus. The genetic and product interaction networks encompass

Table 1: Manual assignment of Gene Ontology evidence codes to enzymatic activities in the metabolic network. For details on the data potentially associated with each evidence code, refer to ref. 30

Gene Ontology evidence code	Evidence code meaning	Number of enzymes with this level of evidence	% of the total number of enzymes in the model
TAS	Traceable author statement	141	17.5
IDA	Inferred from direct assay	57	7.1
IMP	Inferred from mutant phenotype	162	20.0
IPI	Inferred from physical interaction	9	1.1
IGI	Inferred from genetic interaction	13	1.6
IEP	Inferred from expression pattern	130	16.1
ISS	Inferred from sequence similarity	126	15.6
IEA	Inferred from electronic annotation	16	2.0
NAS	Non-traceable author statement	154	19.1

Almost a fifth of entries are fully validated

the control of 715 genes by 89 regulatory molecules, and 160 product interactions involving some 300 proteins.

CONCLUSION

This work describes information standards (in the form of XML DTDs), for the creation and sharing of cellular models, and a progress report on the development of a whole cell model of *E. coli*. The Gene Ontology evidence codes have been used to assess the validity of the data in the model. Incidentally, cell models developed for organisms that are part of the current GO consortium efforts will obviously already have these codes assigned, and hence will not require rechecking.

XML files are human-readable and simple to create

Programmatically, the XML files were designed to facilitate ease of creation and subsequent human reading. Inconsistent nomenclature of metabolites within and between enzyme databases has resulted in the need for a thesaurus. For the sake of completeness, the thesaurus also includes SMILEs (simplified molecular input line entry specification) strings to allow us to make use of the Daylight's computational chemistry tools.[44] The thesaurus is referenced using a client–server model. We envisage the provision of the server source and data files as an aid to programmers and modellers at a later date.

This format facilitates automated updating

Comparable attempts at defining an XML standard for the description of biological networks have been attempted either in parallel, or previously. The most prominent of these are the Systems Biology Markup Language Level 1 (SBML)[45] and Cell Markup Language 1.0 (CellML).[46]

SBML favours the creation of lists (for example ListOfSpecies and ListOfReactions) as XML objects upon which to draw. Consequently, the resulting file is not very human-readable as it contains identifiers to components listed elsewhere. Kinetic information is condensed into one KineticLaw, while our methodology is more flexible and captures more data, incorporating maximum velocity and affinity constants among others. While SBML allows 'annotation' and 'notes' (differentiated as computer-added hidden information and user-added visible information) to be placed anywhere within the file, we have specific locations where this information may be placed, and have various elements to hold different types of annotation such as references and additional identifiers. Importantly, this allows us to introduce comments and references into our file efficiently and automatically, as well as allowing for ease of navigation around the file to check specific comments and annotations.

CellML focuses on encapsulation such that all elements are compartmentalised. Although this encourages reuse of components, it makes initial assembly manually intensive and not so easy to automate. In addition, variables within a compartment must be manually assigned attributes, again making automation difficult. CellML also requires the use of MathML,[47] thereby necessitating the knowledge of one format within another. Currently, work is underway to introduce FieldML[48] into CellML. While these additions are beneficial from the viewpoint of adding object-oriented properties to modelling data, it does mean that the file itself becomes less human-readable.

Presumably, as new data become available the information in SBML and CellML formats will need to be updated. Advantageously, our methodology allows for simple and automated update directly from web resources used, as entries are not interlinked as lists and do not form an encapsulated data set. Both SBML and CellML working groups are currently addressing the interconvertibility issues between their language specifications, and clearly both will be exceptionally useful in facilitating exchange of information between modelling groups by acting as standard model exchange languages. Currently, our data are mapped to either format with relative ease (limited to the some loss of information, dependent on target format chosen for mapping).

Format for whole cell data

The DTDs are now relatively stable and are freely available. However they might be subject to minor modification, to facilitate the incorporation of suggested changes from the scientific community. They are available, together with the gene-regulation classification scheme, at ref. 38.

Cell modelling could reduce the need for animal experimentation

A preliminary whole cell DTD has also been made available. Compliance of XML files with updated DTDs should be relatively straightforward through the use of the appropriate software to handle these files, which will also be made available at the above site.

Model refinement

The creation of a whole cell model is an iterative process, requiring the assembly of the matrices independently, and then establishing the connections between them. It is anticipated that, even with an organism as well studied as *E. coli*, there will be numerous gaps in our knowledge that will lead to deficiencies in the model. It may be possible to circumvent this problem through the use of 'reverse engineering' to predict regulatory interactions.[49]

Model refinement is an iterative process

Uses of models

Metabolic engineering principles (for a review see Yang *et al.*[50]) can be applied to either create *in silico* gene knockouts and metabolic mutations, or to predict the physiological consequences of introducing additional enzymatic pathways.

Whole cell models have a number of potential applications besides prediction of culture conditions to optimise industrial fermentations, and maximise yields of end products or intermediates. From an academic standpoint, whole cell models could improve our understanding not only of cellular networks, but also of other complex systems. Models could be used not only to predict the *in vitro* and *in vivo* physiologies of an organism, but also to improve the knowledge base for the organism in question. Following bacterial cell modelling, eukaryotic cell modelling could revolutionise toxicological and safety assessments, and reduce the need for animal experimentation.

There is still some distance to go in fully validating even the *E. coli* metabolic network model. Any assistance with this task, or suggestions as to how it may be more effectively accomplished, are most welcome.

Acknowledgements

The authors are grateful to Dr Steve Baigent for his suggestions during the revision of this manuscript, as well as for organising this special issue of *Briefings in Bioinformatics*.

References

1. Fleischmann, R. D., Adams, M. D., White, O. *et al.* (1995), 'Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd', *Science*, Vol. 269, pp. 496–512.
2. URL: http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html
3. Giersch, C. (2000), 'Mathematical modelling of metabolism', *Curr. Opin. Plant Biol.*, Vol. 3, pp. 249–253.
4. Gombert, A. K. and Nielsen, J. (2000), 'Mathematical modelling of metabolism', *Curr. Opin. Plant Biotech.*, Vol. 11, pp. 180–186.
5. Schilling, C. H. and Palsson, B. O. (1998), 'The underlying pathway structure of biochemical reaction networks', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 4193–4198.
6. Simpson, T. W., Follstad, B. D. and Stephanopoulos, G. (1999), 'Analysis of the pathway structure of metabolic networks', *J. Biotechnol.*, Vol. 71, pp. 207–223.
7. Schilling, C. H., Letscher, D. and Palsson, B. O. (2000), 'Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective', *J. Theor. Biol.*, Vol. 203, pp. 229–248.
8. Edwards, J. S., Ibarra, R. U. and Palsson, B. O. (2001), '*In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data', *Nature Biotechnol.*, Vol. 19, pp. 125–130.
9. URL: http://e-cell.org/
10. Tomita, M., Hashimoto, K., Takahashi, K. *et al.* (1997), 'E-CELL: software environment for whole-cell simulation', *Bioinformatics*, Vol. 15, pp. 72–84.
11. URL: http://www.nrcam.uchc.edu/
12. Fraser, C. M., Gocayne, J. D., White, O. *et al.* (1995), 'The minimal gene complement of

Mycoplasma genitalium', *Science*, Vol. 270, pp. 397–403.

13. Blattner, F. R., Plunkett, G., Bloch, C. A. *et al.* (1997), 'The complete genome sequence of *Escherichia coli* K-12', *Science*, Vol. 277, pp. 1453–1474.
14. URL: http://www.expasy.ch/sprot/sprot-top.html
15. URL: http://wit.mcs.anl.gov/
16. URL: http://www.genome.ad.jp/kegg/
17. URL: http://www.empproject.com
18. URL: http://www.expasy.ch/enzyme/
19. URL: http://www.brenda.uni-koeln.de/
20. URL: http://genprotec.mbl.edu/
21. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S. *et al.* (2001), 'RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12', *Nucleic Acids Res.*, Vol. 29, pp. 72–74.
22. URL: http://web.bham.ac.uk/bcm4ght6/genome.html
23. Mahmoudian, M. (2000), 'Biocatalytic production of chiral pharmaceutical intermediates', *Biocatal. Biotransform.*, Vol. 18, pp. 105–118.
24. Caldwell, S. R., Varghese, J. and Puri, N. K. (1996), 'Large scale purification process for recombinant NS1-OspA as a candidate vaccine for Lyme disease', *Bioseparation*, Vol. 6, pp. 115–123.
25. Madurawe, R. D., Chase, T. E., Tsao, E. I. and Bentley, W. E. (2000), 'A recombinant lipoprotein antigen against Lyme disease expressed in *E. coli*: fermentor operating strategies for improved yield', *Biotechnol. Prog.*, Vol. 16, pp. 571–576.
26. Steinberg, F. M. and Raso, J. (1998), 'Biotech pharmaceuticals and biotherapy: an overview', *J. Pharm. Pharm. Sci.*, Vol. 1, pp 48–59.
27. Kane, J. F. (1993), 'Environmental assessment of recombinant DNA fermentations', *J. Indus. Microbiol.*, Vol. 11, pp. 205–208.
28. Pennisi, E. (1999), 'Keeping genome databases clean and up to date', Science, Vol. 286, pp. 447–450.
29. Stevens, R., Goble, C. A. and Bechhofer, S. (2000), 'Ontology-based knowledge representation for bioinformatics', *Briefings Bioinformatics*, Vol. 1, pp. 398–414.
30. URL: http://www.geneontology.org/GO.evidence.html
31. URL: http://www.w3.org/TR/REC-xml
32. URL: http://www.w3.org/Math/
33. URL: http://www.xml-cml.org/
34. Achard, F., Vaysseix, G. and Barillot, E. (2001), 'XML, bioinformatics and data integration', *Bioinformatics*, Vol. 17, pp. 115–125.
35. URL: http://www.w3.org/XML/Schema
36. Yuh, C. H., Bolouri, H. and Davidson, E. H. (1998), 'Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene', *Science*, Vol. 279, pp. 1896–1902.
37. Serov, V. N., Spirov, A. V. and Samsonova, M. G. (1998), 'Graphical interface to the genetic network database GeNet', *Bioinformatics*, Vol. 14, pp. 546–547.
38. URL: ftp://ftp.cds.caltech.edu/pub/goryanin/paper
39. Goryanin, I., Hodgman, T. C. and Selkov, E. (1999), 'Mathematical simulation and analysis of cellular metabolism and regulation', *Bioinformatics*, Vol. 15, pp. 749–758.
40. Varma, A. and Palsson, B. O. (1994), 'Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild type *Escherichia coli* W3110', *Appl. Environ. Microbiol.*, Vol. 60, pp. 3724–3731.
41. Marin-Sanguino, A. and Torres, N. V. (2000), 'Optimization of tryptophan production in bacteria. Design of a strategy for genetic manipulation of the tryptophan operon for tryptophan flux maximization', *Biotechnol. Prog.*, Vol. 16, pp.133–145.
42. Schuster, S., Dandekar, T. and Fell, D. A. (1999), 'Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering', *Trends Biotechnol.*, Vol. 17, pp. 53–60.
43. URL: http://wit.integratedgenomics.com/WIT2/CGI/org.cgi?where=specific_statistics&type=noseq &org=EC
44. URL: http://www.daylight.com
45. URL: http://www.cds.caltech.edu/erato/sbml-level-1/sbml-level-1.html
46. URL: http://www.cellml.org/public/specification/cellml_specification.html
47. URL: http://www.w3.org/TR/MathML2/
48. URL: http://www.physiome.org.nz/sites/physiome/fieldml/pages/index.html
49. Tavazoie, S., HugURL: hes, J. D., Campbell, M. J., Cho. R. J. and Church, G. M. (1999), 'Systematic determination of genetic network architecture', *Nature Genetics*, Vol. 22, pp. 281–285.
50. Yang, Y.-T., Bennett, G. N. and San, K.-Y. (1998), 'Genetic and metabolic engineering', *Electronic J. Biotechnol.*, Vol. 1(3), article 3. Available at: http://www.ejb.org/content/vol1/issue3/full/3/