

Loosely coupled bioinformatics workflows for systems biology via Taverna

Douglas Kell

School of Chemistry, and The Manchester Interdisciplinary Biocentre,
The University of Manchester, MANCHESTER M1 7DN, U.K.

dbk@manchester.ac.uk

<http://dbkgroup.org/>

<http://www.mib.ac.uk> www.mcisb.org

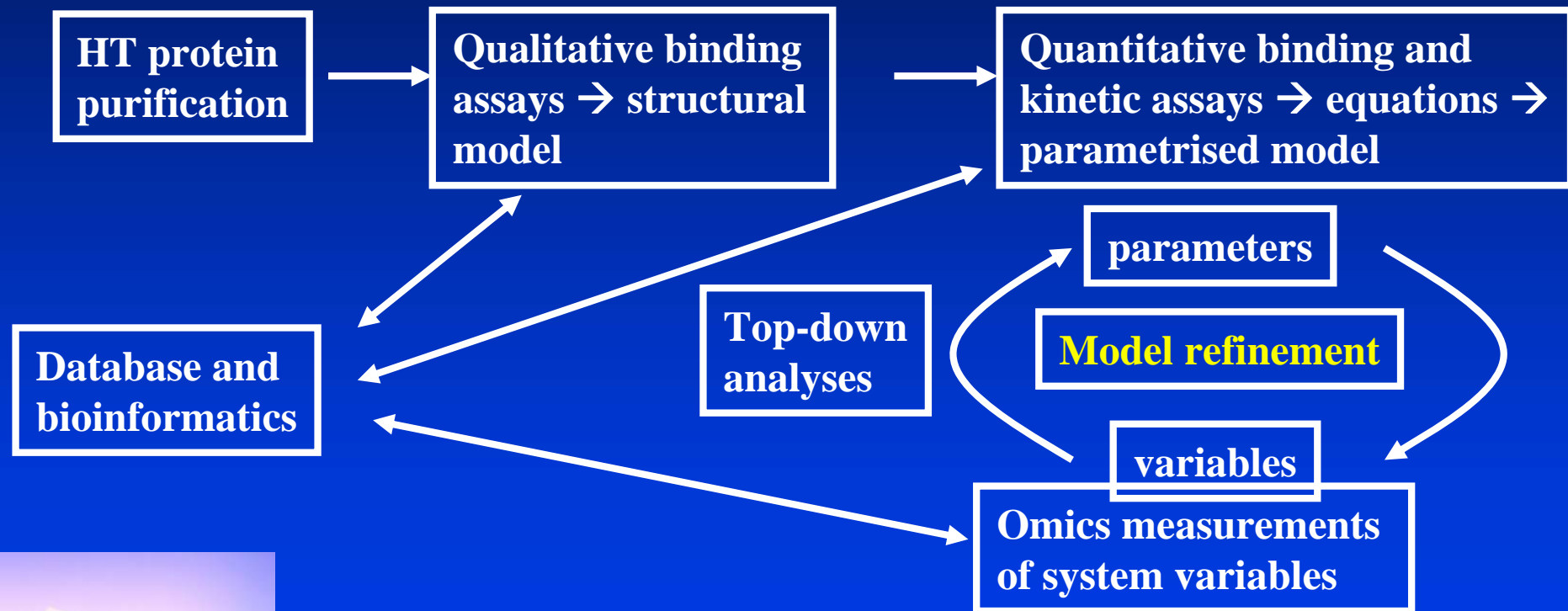


MANCHESTER
1824

The University of Manchester



Basic 'bottom-up'-driven Systems Biology pipeline at MCISB



VISUALISE

Layouts and views

SBGN

Overlays, dynamics

create

edit

Literature mining

BIOCHEMICAL MODEL (assumed to be in SBML)

Store in dB

Model merging: (not)
LEGO blocks

Compare with
other models

Cheminformatic
analyses

**THERE ARE MANY POSSIBLE THINGS THAT ONE
MIGHT DO WITH THIS REPRESENTATION, AND
THESE ACTIONS CAN BE SEEN AS MODULES**

Run, analyse
(sensitivities, etc)

Compare with and fit to real
data (parameters and
variables) with constraints

Integrate various
levels

Store results of
manipulations

How to deal with fitting,
including as $f(\text{global parameters like pH})$

LINK WORKFLOWS

Soaplab, Taverna,

Web services, etc.

Network Motif
discovery

Optimal DoE for
Sys Identification,
incl identifiability

Automatic characterisation
of parameter space and
constraint checking

The Data Management Infrastructure of the Manchester Centre for Integrated Systems Biology

Norman Paton

University of Manchester



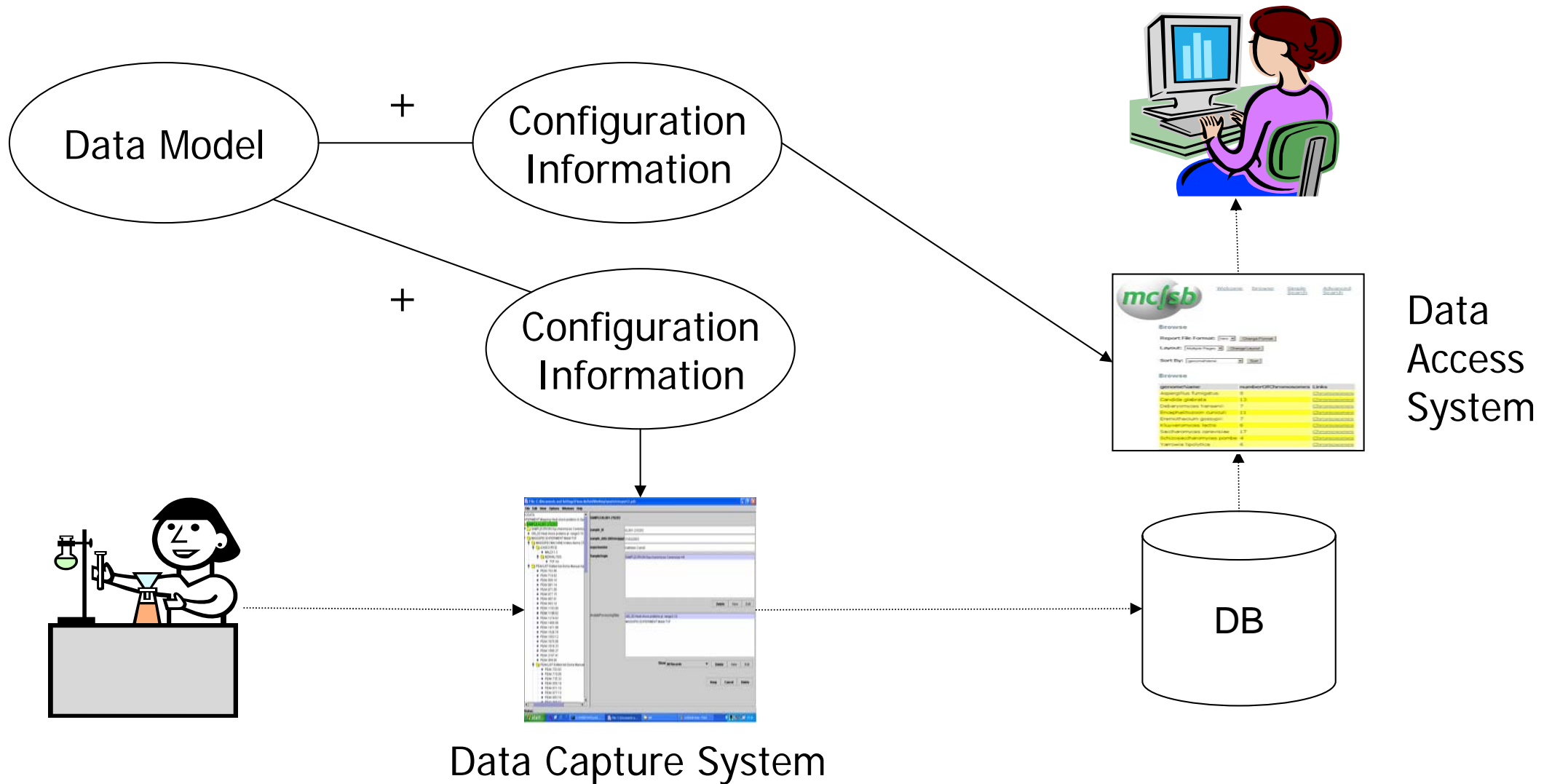
Problem Statement

- **There is a requirement to:**
 - Store experimental data of many different kinds for integration and analysis.
 - Selectively make the data available within the MCISB, to partner sites and for public access.
 - Support programmatic as well as interactive access to experimental data and models.
 - Provide consistent interfaces, to reduce learning times and development costs.

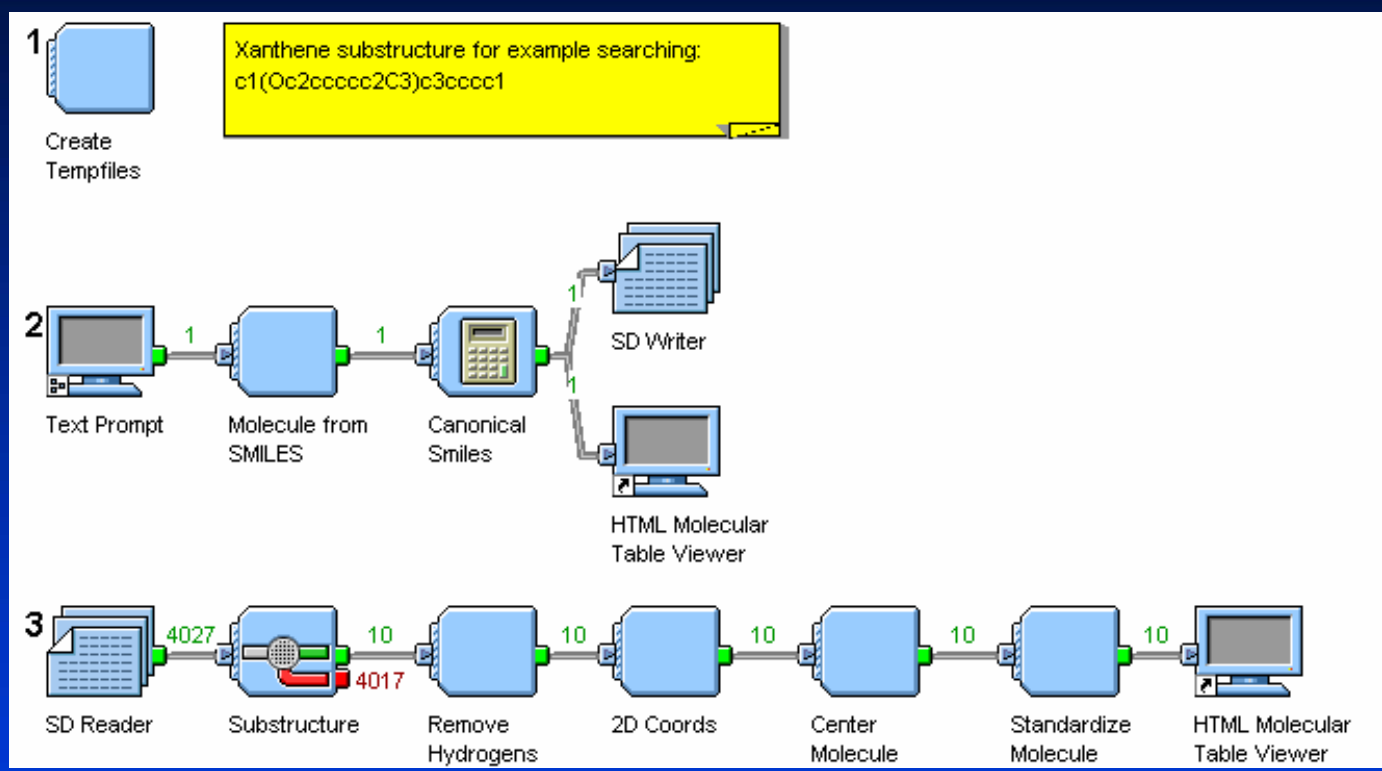
Capabilities

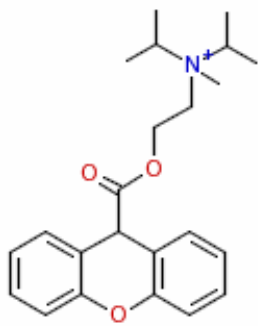
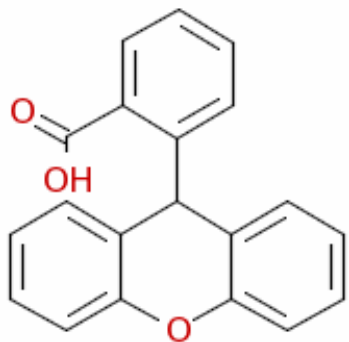
- **We require software to support:**
 - Data capture: Pedro.
 - Data access: Pierre.
 - Integration of data and analyses: Taverna.

Generating Interfaces from Models

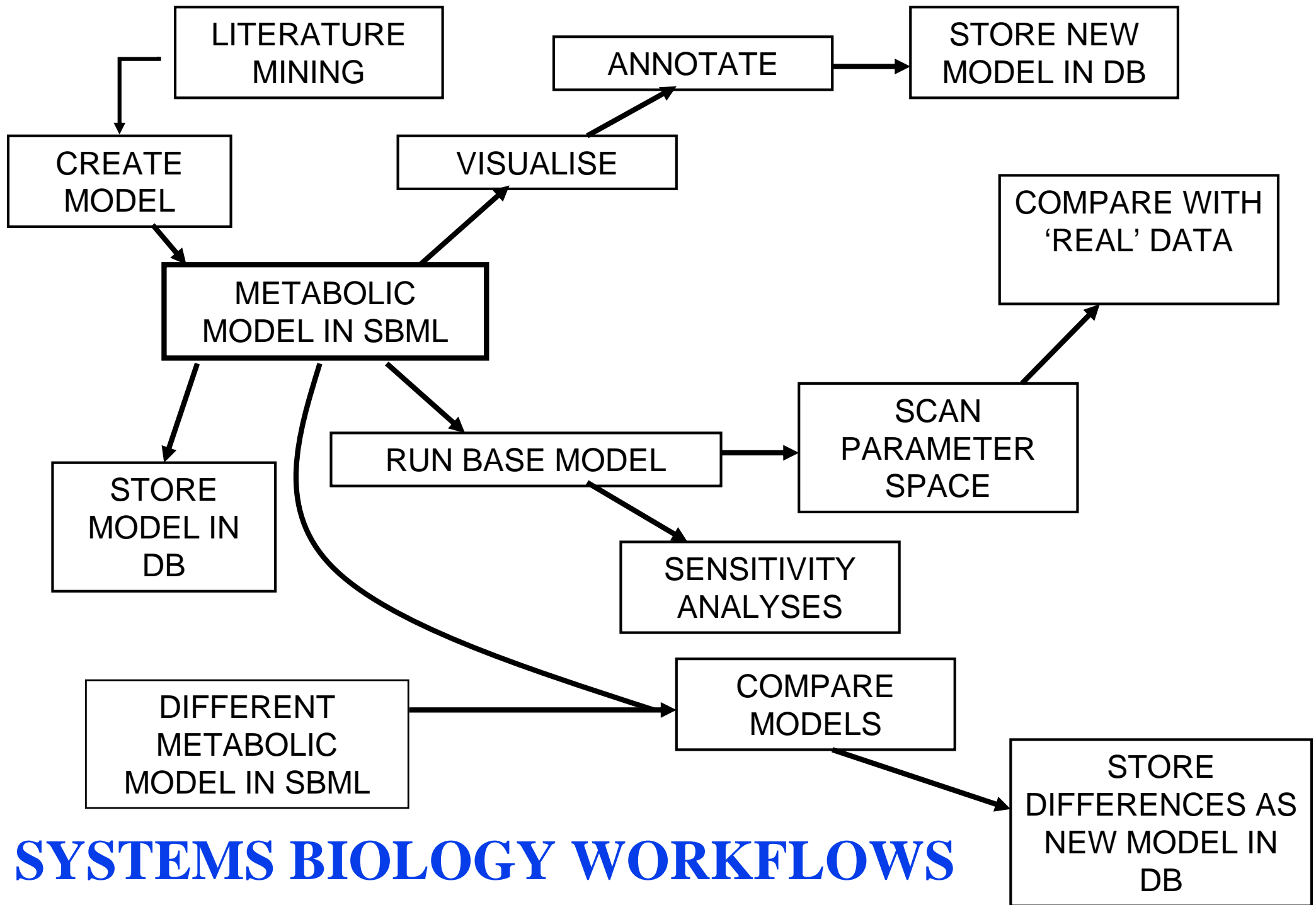


Pipeline Pilot workflow



	Br^-	Propanteline bromide
		28311
		Chiral

etc...



SYSTEMS BIOLOGY WORKFLOWS

Scientists

Decoupled suppliers & consumers

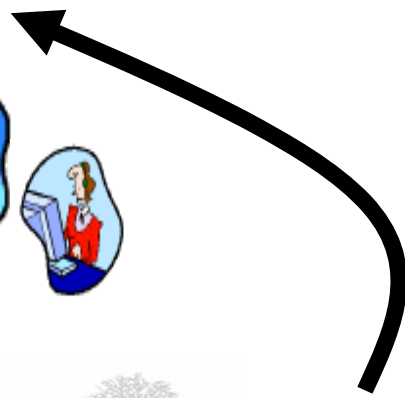
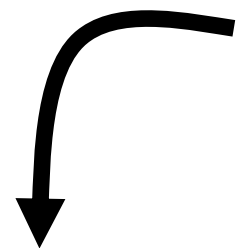


VC 2005 8th Novem



Science

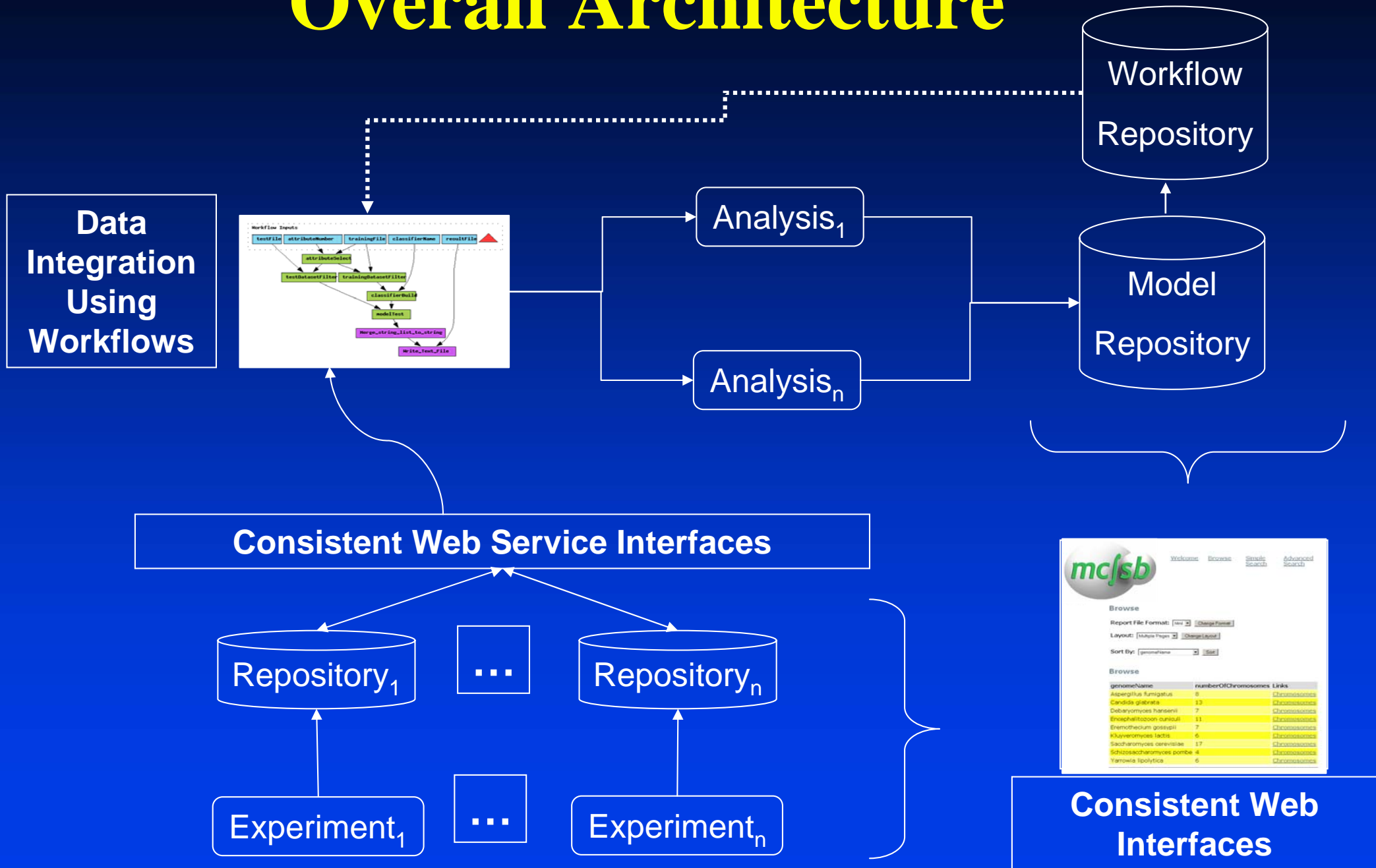
DDBJ DNA Data Bank of Japan
EBI-EBI European Bioinformatics Institute
bio::mart
SoapLab
wellcome trust sanger institute
NCBI
cellML
SML
MOUNT SINAI HOSPITAL
PDBj xp Protein Data Bank Japan Experimental Version
PathPort The Pathogen Portal Web Project
emboss
moby
LION
CABIO



‘Warehouse’ vs distributed workflows

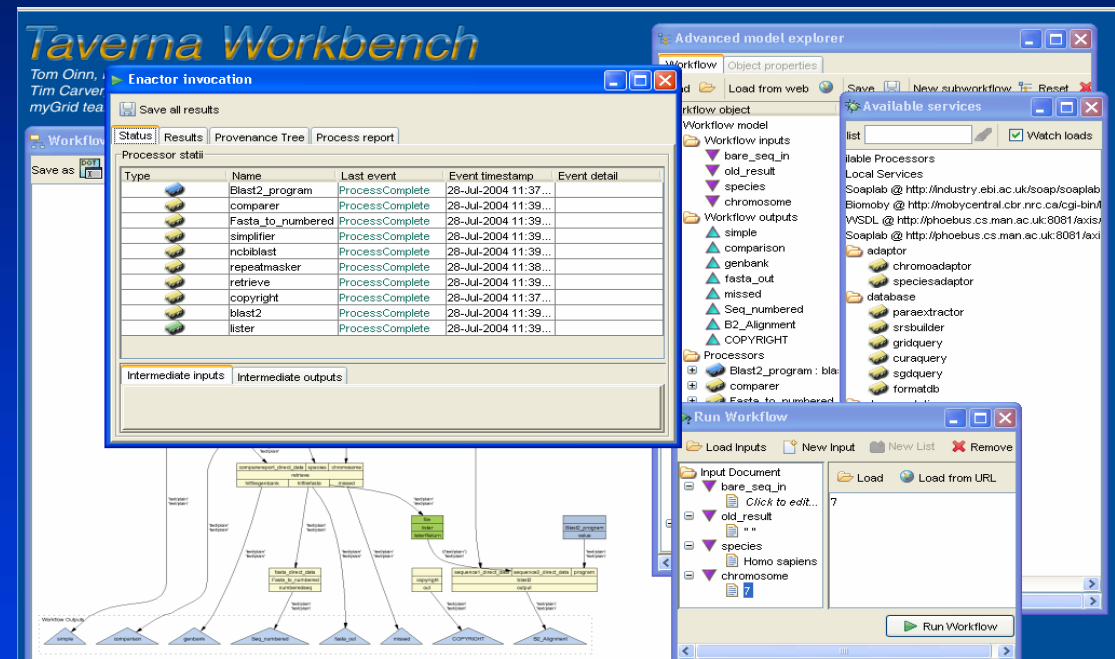
- **Different ‘modules’ developed in different labs can reside on different computers anywhere, and expose themselves as Web Services**
- **Labs can then specialise in what they are best at**
- **All that is then needed is an environment for enacting bioinformatic workflows by coupling together these service-oriented architectures**
- **One such is Taverna**
- **This is arguably the best way to combine metabolomic SBML models with metabolomic data, and is what we plan to do at MCISB**

Overall Architecture



Taverna Workflow Environment

- Workflow environment for authoring scientific workflows.
- Developed by myGrid e-Science Pilot project.
- Downloads: over 1000 a month during 2006.



<http://taverna.sourceforge.net/>

Taverna (sits on myGrid)

www.mygrid.org.uk

www.taverna.sf.net



BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3045–3054
doi:10.1093/bioinformatics/bth361



Taverna: a tool for the composition and enactment of bioinformatics workflows

Tom Oinn¹, Matthew Addis², Justin Ferris², Darren Marvin²,
Martin Senger¹, Mark Greenwood³, Tim Carver⁴, Kevin Glover⁵,
Matthew R. Pocock⁶, Anil Wipat⁶ and Peter Li^{6,*}

myExperiment.org



Taverna Workbench

Advanced model explorer

Workflow: Metadata for 'GetDiseaseGeneIDs'

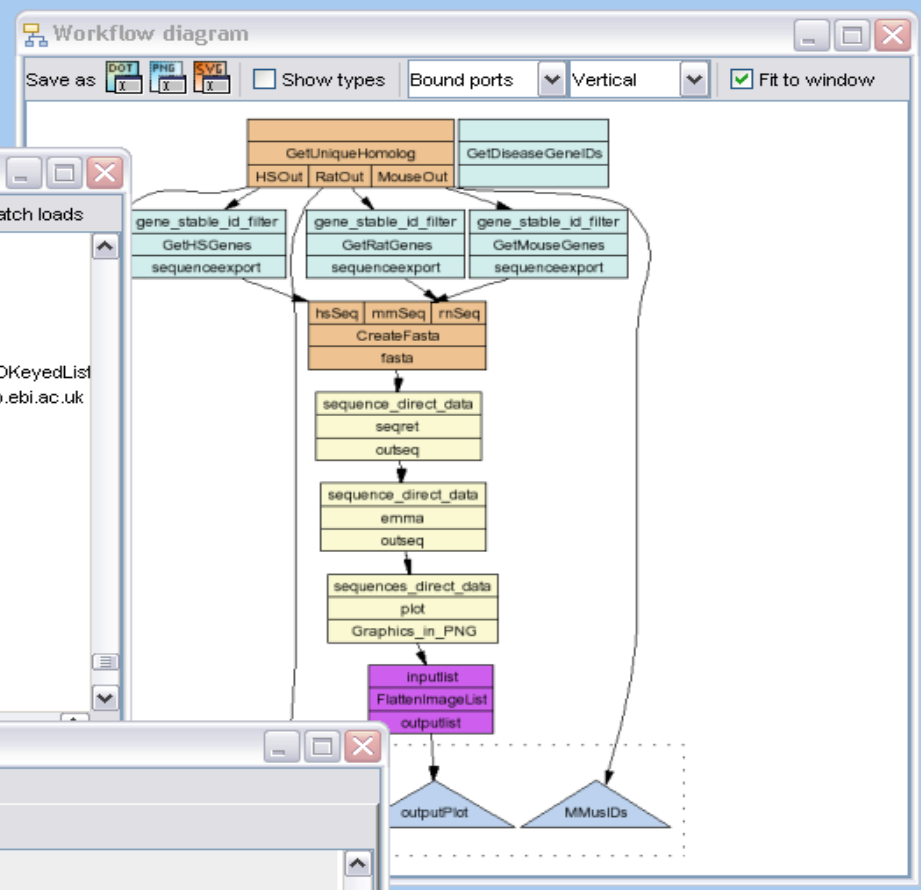
Load Load from web Save New subworkflow Offline Reset

Workflow object	Retries	Delay	Backoff	Threads	Critical
Processors					
GetUniqueHomolog	0	0	1	1	<input type="checkbox"/>
GetMouseGenes	0	0	1	1	<input type="checkbox"/>
GetHSGenes	0	0	1	1	<input type="checkbox"/>
GetRatGenes	0	0	1	1	<input type="checkbox"/>
CreateFasta	0	0	1	1	<input type="checkbox"/>
hsSeq					
mmSeq					
rnSeq					
fasta					
GetDiseaseGeneIDs	0	0	1	1	<input type="checkbox"/>
chr_name_filter					
sequenceexport					
FlattenImageList	0	0	1	1	<input type="checkbox"/>
seqret	0	0	1	5	<input type="checkbox"/>
emma	0	0	1	5	<input type="checkbox"/>
plot	0	0	1	5	<input type="checkbox"/>
Data links					
GetUniqueHomolog:HSOut->GetH					

Available services

Search list Watch loads

- GetDomainsFromGWWithEvalve
- GetAccFromRetiredGi
- ProteinReportSetDescription
- GetFastaKeyedList
- RedundantGroupKeyedList
- GetFastaFromRedundantGroupIDKeyedList
- Biomart ensembl_mart_22_1@martdb.ebi.ac.uk
 - frubripes_gene_ensembl
 - hsapiens_gene_est
 - cbriggsae_gene_est
 - rnorvegicus_gene_est
 - drerio_gene_ensembl
 - ggallus_gene_ensembl
 - celegans_gene_ensembl
 - rnorvegicus_gene_ensembl
 - agambiae_gene_est
 - drerio_gene_est
 - ggallus_gene_est
 - cbriggsae_gene_ensembl



Enactor invocation

Save as XML Save to disk Save to disk as website

Status Results Process report

MMusIDs HSapiDs RNorIDs outputPlot

/saraty/emboss-interfaces/a/unkn

Species	Sequence
Mouse	G G C A C C C T C A T T T C C T G C F A C C C C
Rat	- - - - G A C T C G T G C G C C A G C A C C C T T
Human	- G C T A T T T T A T T T T T A G T F G A C F A
Mouse	G G F C A T T C C C T A G G C C F C T C - - - B T G
Rat	G G C T A C T T G G A A G G C A C T T C C C G G F G
Human	G G T A C G T A G T G A F G F T F C T B T A C - B T G
Mouse	G A G C T C G G G T C C T A C C T C C T C E G C C A
Rat	G T G C A C C G G T T A A A C C C E G G C A G T F A G
Human	T A G C A C A C C A C C A C T C A C F A C A C A G
Mouse	T C
Rat	C T
Human	T C
Mouse	G G
Rat	C C C
Human	A C C

Advanced model explorer

Workflow: Remote resource usage

Save HTML description

Resource usage report

This display shows the various external resources used by the workflow. It does not show resources such as local operations or string operations performed by the enactment engine. Services are categorized by name of the instance of each service shown to the right.

Resources on martdb.ebi.ac.uk, 4 instances.		
Biomart	Dataset Name	Proc
	mmusculus_gene_ensembl	GetM
Biomart	Dataset Name	Proc

Configuring query for GetHSGenes

Attributes Filters

Features Structures Sequences SNPs

Sequences

Type of sequence to export: REGION: GENE: PROTEIN:

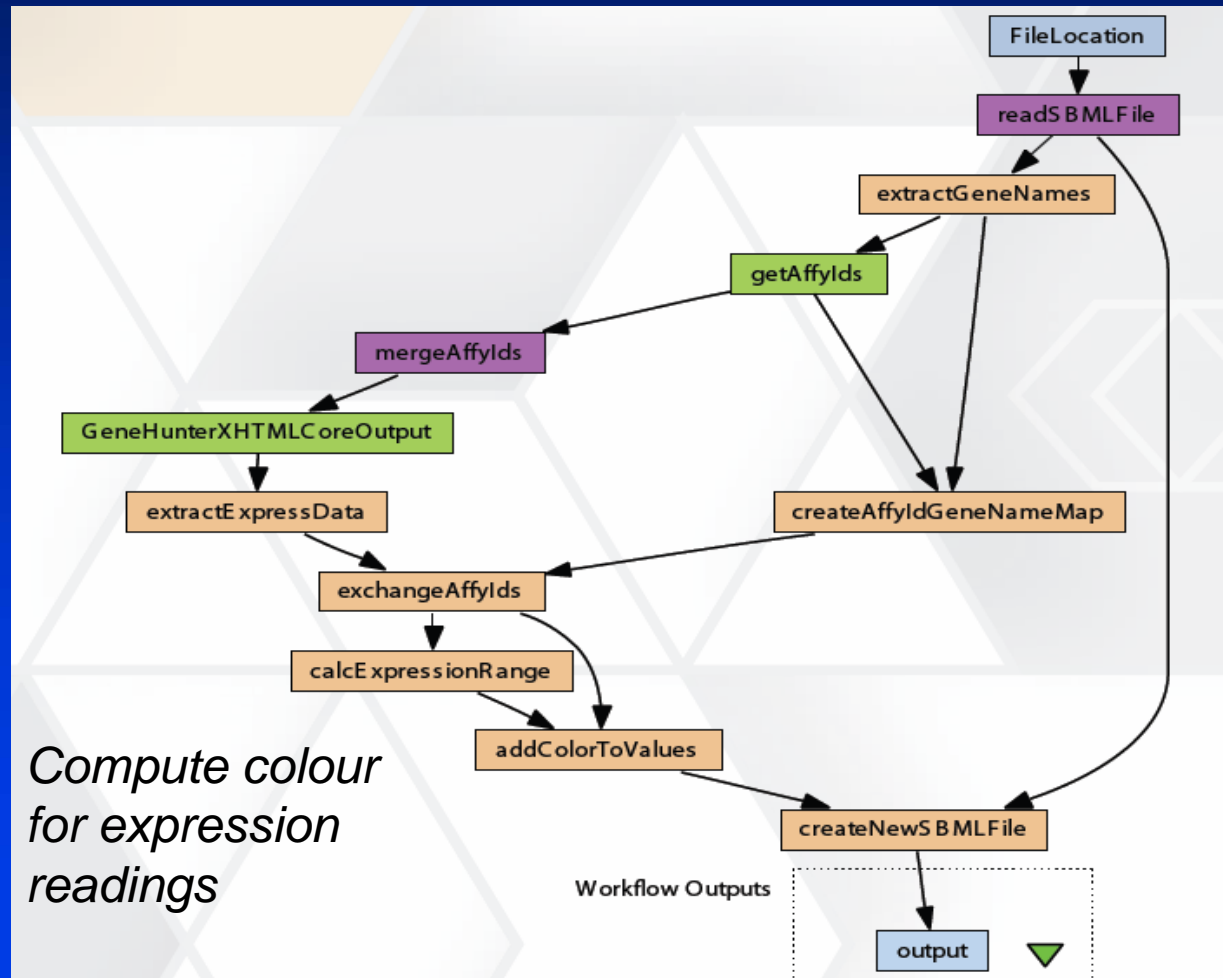
Sequence export options.

Type of sequence to fetch

Taverna Workflow Workbench

Relating Models to Expression

*Query maxd
transcriptome
database using
gene names*

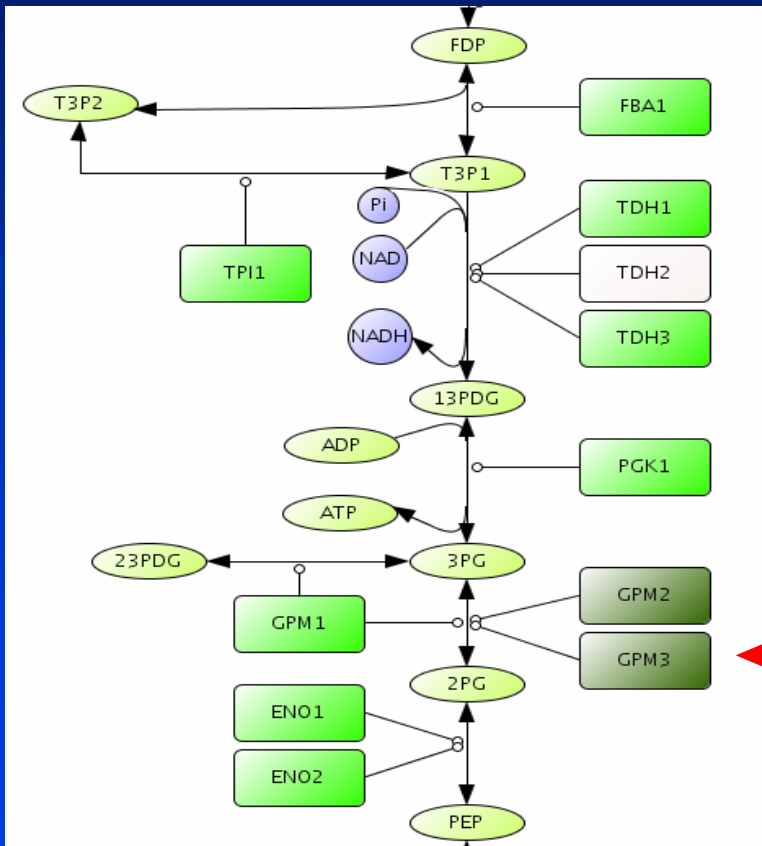


*Compute colour
for expression
readings*

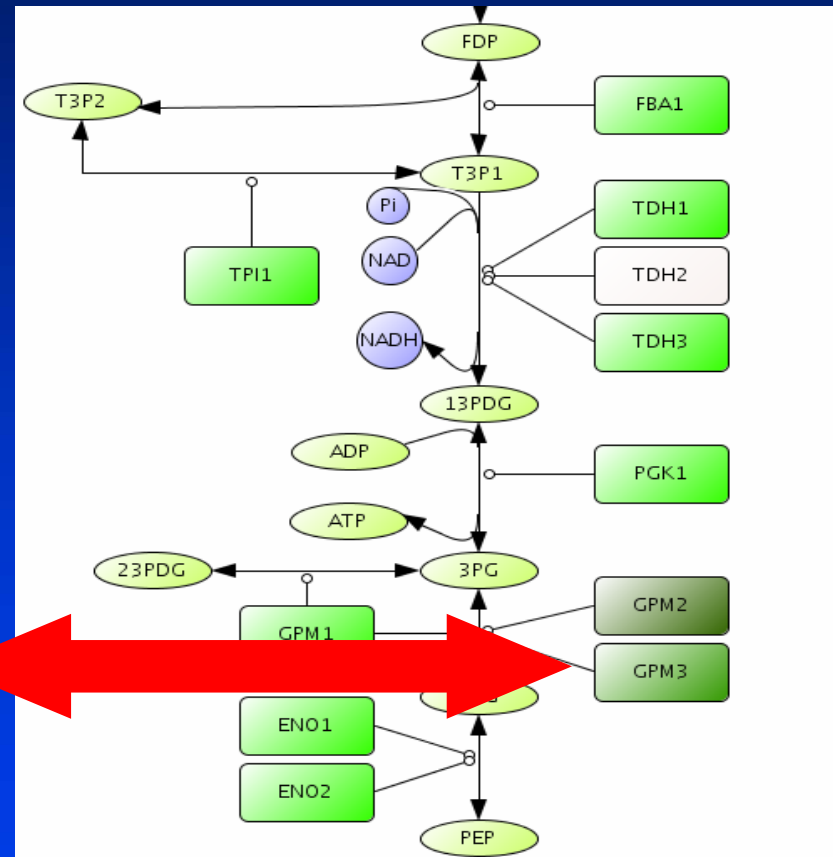
*Read gene
names of
enzymes from
SBML model*

*Create new
SBMLmodel*

Visualise Models Using Cell Designer

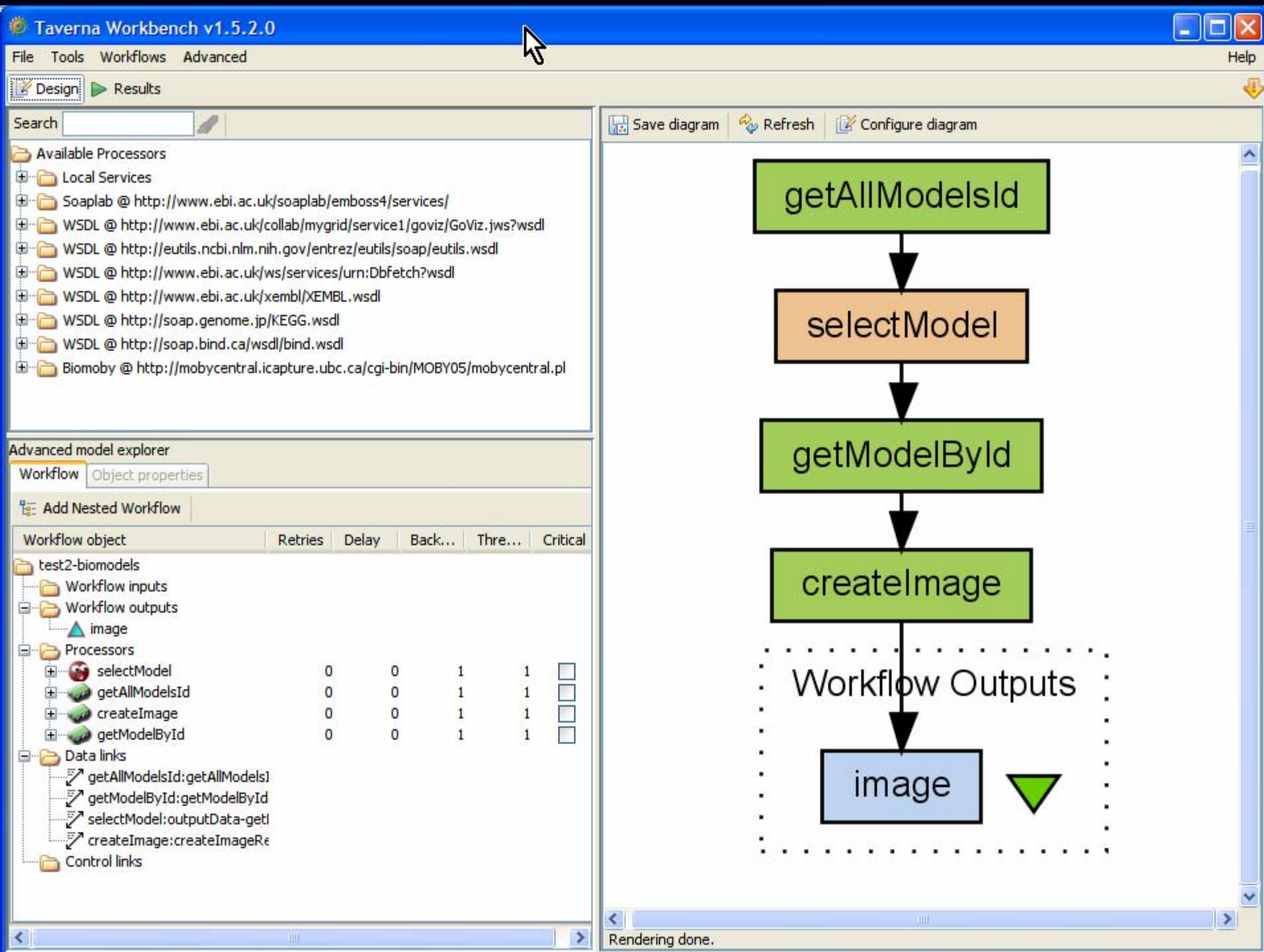


JC_C-0.07-1_Measurement

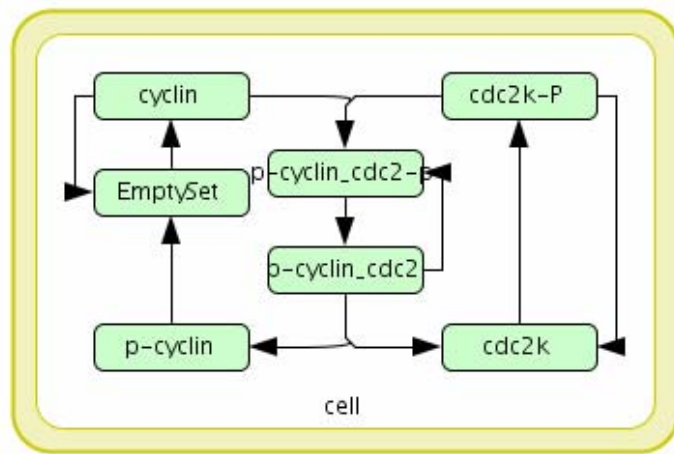


JC_N-0.07-1_Measurement





- List
- urn:lsid:net.sf.taverna:dataCollection:a223db7f-d27e-4951-9fed-5196f66ba6f9
- application/octet-stream,image/png
- urn:lsid:net.sf.taverna:dataItem:f3dbf944-a77e-4c7b-be25-9619693b9dd0



Issues

- **Legal SBML may not be decent; much more control is needed to effect useful interoperability**
- **Naming conventions are a big issue**
- **Much better semantic markup needed to make comparisons easy**
- **Ontologies remain problematic**
- **We cannot presently visualise large networks; these break the existing software - major effort needed here**

Naming issues....

- One involves the fact that aspartate and aspartic acid are not the same molecule
- For many purposes, e.g. FBA, this does matter a lot
- One solution: calculate or look up pKa values and include pH of each compartment explicitly in the formal part of the SBML model – **how?**

Representing parameter uncertainty

- Present SBML stores parameters as crisp values, but not any attendant measures of uncertainty in them \wedge vs $|$
- Various kinds of modelling, especially but not only Bayesian ones, really need access to such measures of uncertainty
- An important set of questions for future SBML annotation relates to these measures of uncertainty and how they should optimally be encoded
- Please let us develop these! Some are obvious (Gaussian with a mid-value and a non-negative range) and can be expressed as equations just as in the models themselves. Others may be less so. We also need annotations of the attendant evidence that led to these uncertainties
- (Same is true for databases of experimental parameters....)

Thanks to....



Loosely coupled bioinformatics workflows for systems biology via Taverna

Douglas Kell

School of Chemistry, and The Manchester Interdisciplinary Biocentre,
The University of Manchester, MANCHESTER M1 7DN, U.K.

dbk@manchester.ac.uk

<http://dbkgroup.org/>

<http://www.mib.ac.uk> www.mcisb.org



MANCHESTER
1824

The University of Manchester

