

2009-01-22

Jukka Matilainen
jukka.matilainen@medicel.com

State Data

Relating Quantitative Data to Pathway Network Models

Presentation contents

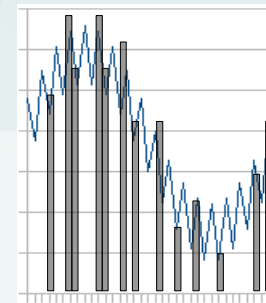
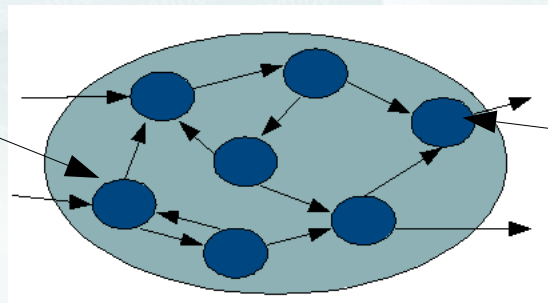
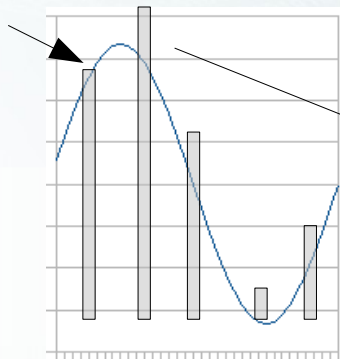
- **Why** do we need state data?
- **What** exactly is represented by state data?
 - state of a system
 - measurement results
 - simulation results
- **Where** does state data come from?
 - External sources
 - Internal sources
- **Summary**

Part 1 - Why do we need State data?

Background: State of a system

- "State data" :
 - observations/predictions (quantitative values) for quantities related to components of the system under study at
 - reflects the state of a system at a given point of time
- State vector: multiple variables, only some of which may be "visible" (known/measurable)

State data

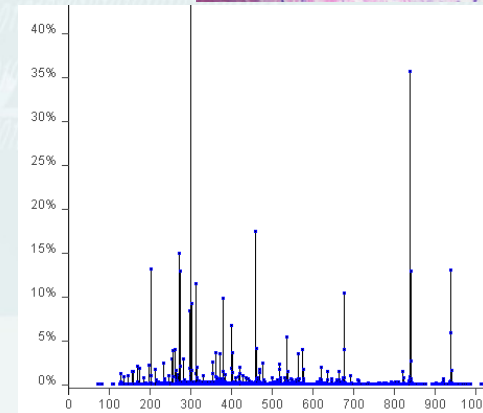
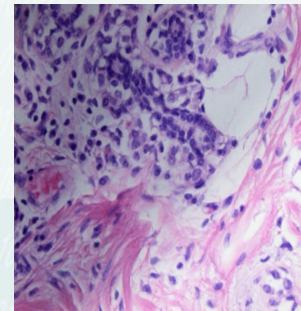
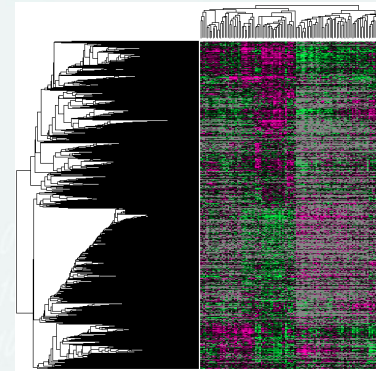


State data

Part 2 - What is state data in Integrator?

State data in biological context

- measurement data (observations of components), such as
 - gene (mRNA) expression levels
 - protein concentration
 - protein, glycopeptide, peptide quantification
 - particle number/concentration in microscopy image(s)
 - etc.
- simulation data
 - RNA/protein/compound concentration
 - Component absence / presence information



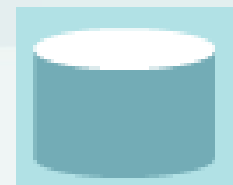
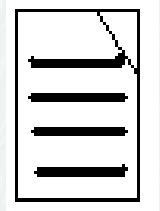
State Data in Integrator: Example

- Informally:

“the measured concentration (in mmol/L) of PKC within the *purkinje cell cytosol* at 2008.01.12 10:23:40 taken from sa001 was 1.2838”
- Formally:
 - 1) observed **variable**: “concentration”
 - 2) **unit** representation: “mmol/L”
 - 3) observed **component**: “PKC”
 - 4) *Optionally*: in which **model**: -
 - 5) In which biological **system**: “purkinje cell cytosol”
 - 6) At what **time**: “2008.01.12 10:23:40”
 - 7) The **sample** (optionally): “sa001”
- The observed **value**: “1.2838”

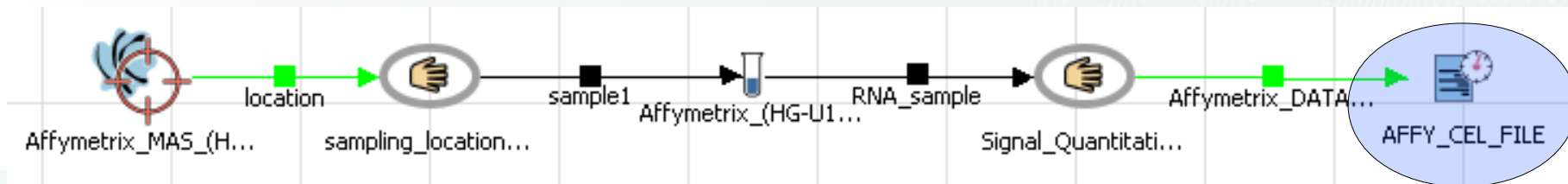
State Data: data entities vs DB tables

- Both respond to different needs:
- 1) Data Entities
 - Processed by tools within Workflow
 - When potentially long calculus operations are needed
- 2) Database tables
 - Used for temporary storage, when specific *data sets* are needed for multidimensional OLAP -style visualization
 - Used by Query, State, Pathway (GUI Applications)



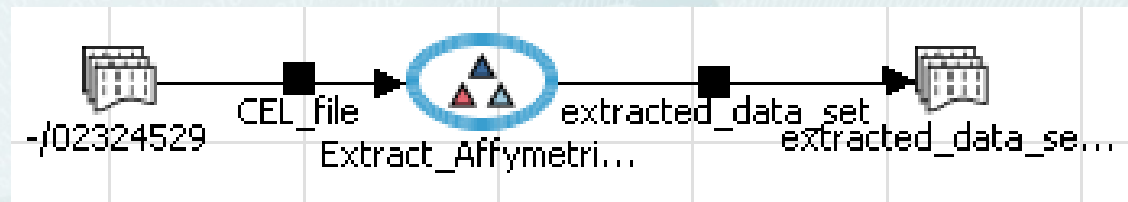
1) State data in data entities

- State data (measurement data) is imported into **Experiment**, at the end of the wet-lab process



- Raw data is first extracted from raw data entities with

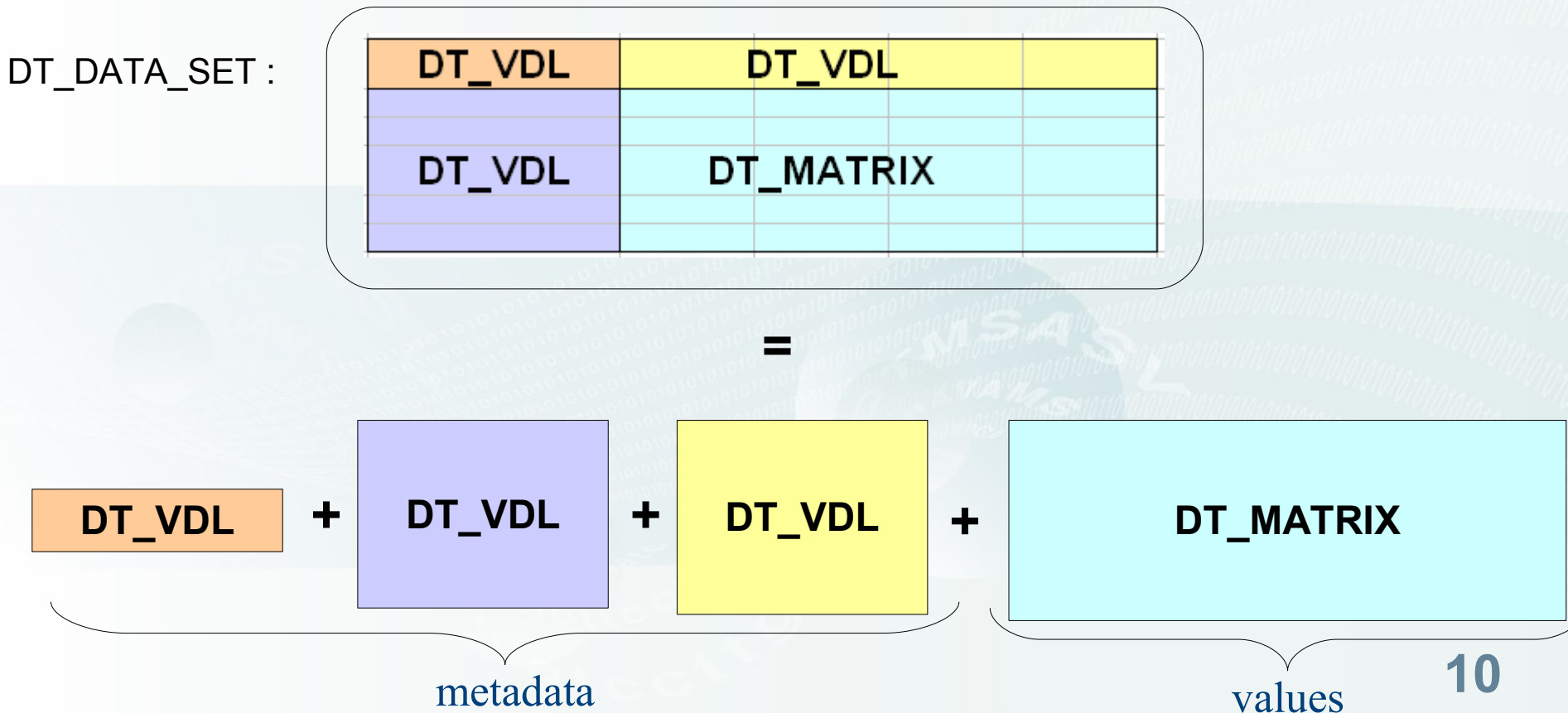
Workflow



- Results of extraction are typically stored in `DT_DATA_SET` format, where the essential metadata is attached with the data values

1) State data in data entities - DT_DATA_SET

Recapitulation: DT_DATA_SET - a composite data type



1) State data in data entities - DT_DATA_SET

An example of extracted Affymetrix state data in DT_DATA_SET format :

V[expression] U[-] L[HS liver cell cytoplasm]	Sa[sa_001] Ts[2005.10.02 12:00:05]	Sa[sa_002] Ts[2005.10.02 15:50:15]	Sa[sa_003] Ts[2006.10.06 11:00:05]	Sa[sa_004] Ts[2007.02.02 09:18:28]
Tr[ENST00000235]	1,65	7,12	5,12	2,65
Tr[ENST00000237]	2,77	9,23	7,12	0,05
Tr[ENST00000239]	5,99	6,23	0	99,22
Tr[ENST00000250]	10,23	8,8	9	100,87
Tr[ENST00000288]	29,02	1,66	7,55	18,77

=

DT_VDL

+

DT_VDL

+

DT_VDL

+

DT_MATRIX

1) State data in data entities - DT_DATA_SET

An example of extracted Affymetrix state data in DT_DATA_SET format :

V[expression] U[-] L[HS liver cell cytoplasm]	Sa[sa_001] Ts[2005.10.02 12:00:05]	Sa[sa_002] Ts[2005.10.02 15:50:15]	Sa[sa_003] Ts[2006.10.06 11:00:05]	Sa[sa_004] Ts[2007.02.02 09:18:28]
Tr[ENST00000235]	1,65	7,12	5,12	2,65
Tr[ENST00000237]	2,77	9,23	7,12	0,05
Tr[ENST00000239]	5,99	6,23	0	99,22
Tr[ENST00000250]	10,23	8,8	9	100,87
Tr[ENST00000288]	29,02	1,66	7,55	18,77

=>

8,8 = V[expression]U[-]L[HS_liver_cell_cytoplasm]Tr[ENST00000250]
Sa[sa_002]Ts[2005.10.02 15:50:15]

VDL – “Variable Description Language”

- Descriptive metadata formatted in key - value pairs
- Keys refer to *database tables*
- Values refer to *database rows*

V[expression]U[-]L[HS_liver_cell_cytoplasm]

DT_VDL

V[expression] U[-] L[HS_liver_cell_cytoplasm]	Sa[sa_001] Ts[2005.10.02 12:00:05]	Sa[sa_002] Ts[2005.10.02 15:50:15]	Sa[sa_003] Ts[2006.10.06 11:00:05]	Sa[sa_004] Ts[2007.02.02 09:18:28]
Tr[ENST00000235]	1,65	7,12	5,12	2,65
Tr[ENST00000237]	2,77	9,23	7,12	0,05
Tr[ENST00000239]	5,99	6,23	0	99,22
Tr[ENST00000250]	10,23	8,8	9	100,87
Tr[ENST00000288]	29,02	1,66	7,55	18,77

Variable : “expression”

Unit : “_”

Location : “HS_liver_cell_cytoplasm”

VDL – “Variable Description Language”

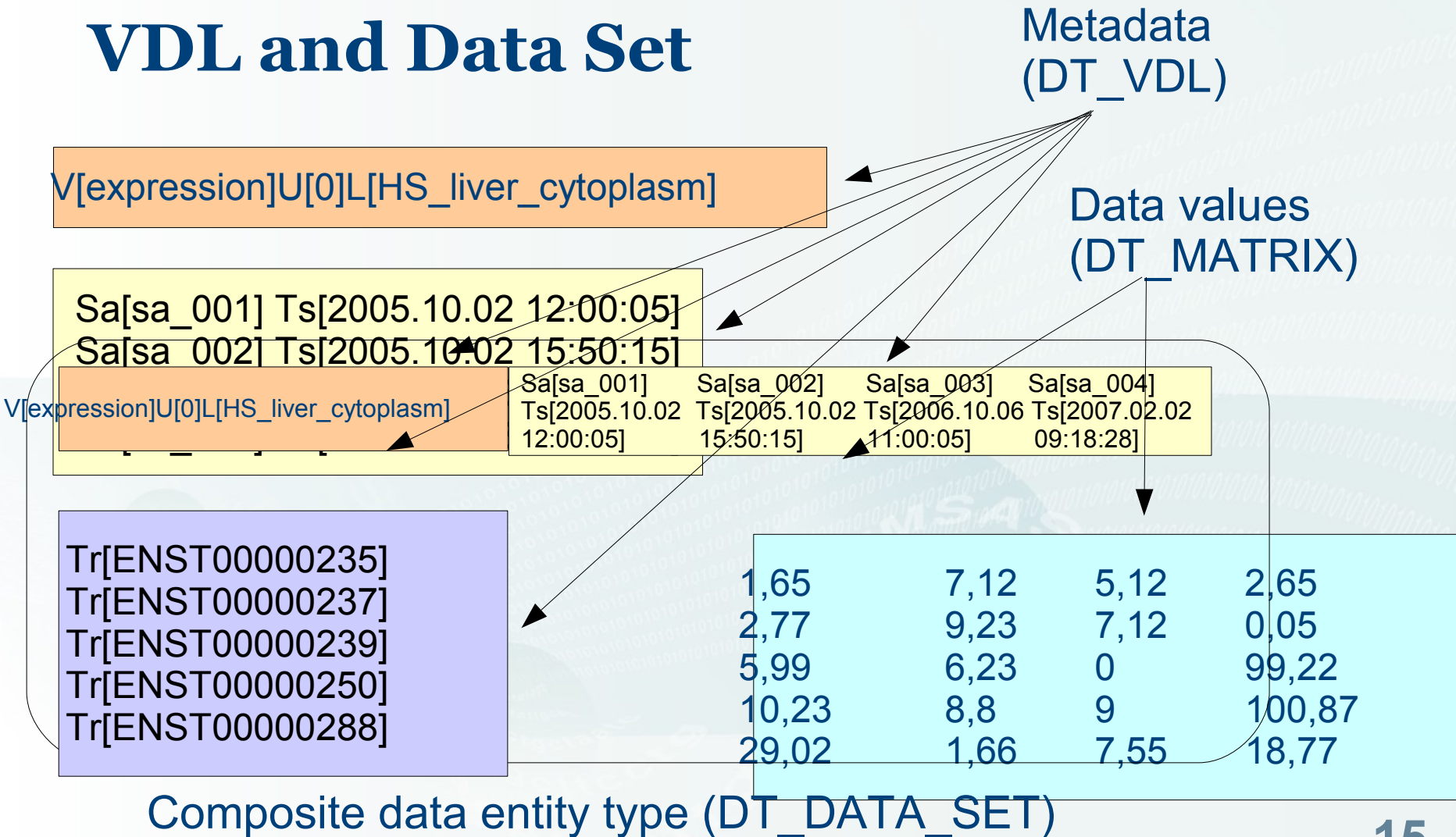
```
Tr[ENST00000235]  
Tr[ENST00000237]  
Tr[ENST00000239]  
Tr[ENST00000250]  
Tr[ENST00000288]
```

```
Sa[sa_001] Ts[2005.10.02 12:00:05]  
Sa[sa_002] Ts[2005.10.02 15:50:15]  
Sa[sa_003] Ts[2006.10.06 11:00:05]  
Sa[sa_004] Ts[2007.02.02 09:18:28]
```

- Similarly for other keys:
 - “Tr” -> Transcript
 - “Sa” -> Sample
 - “Ts” -> Timestamp
 - ...
- There exists a separate key for each supported component

State data in data entities

VDL and Data Set

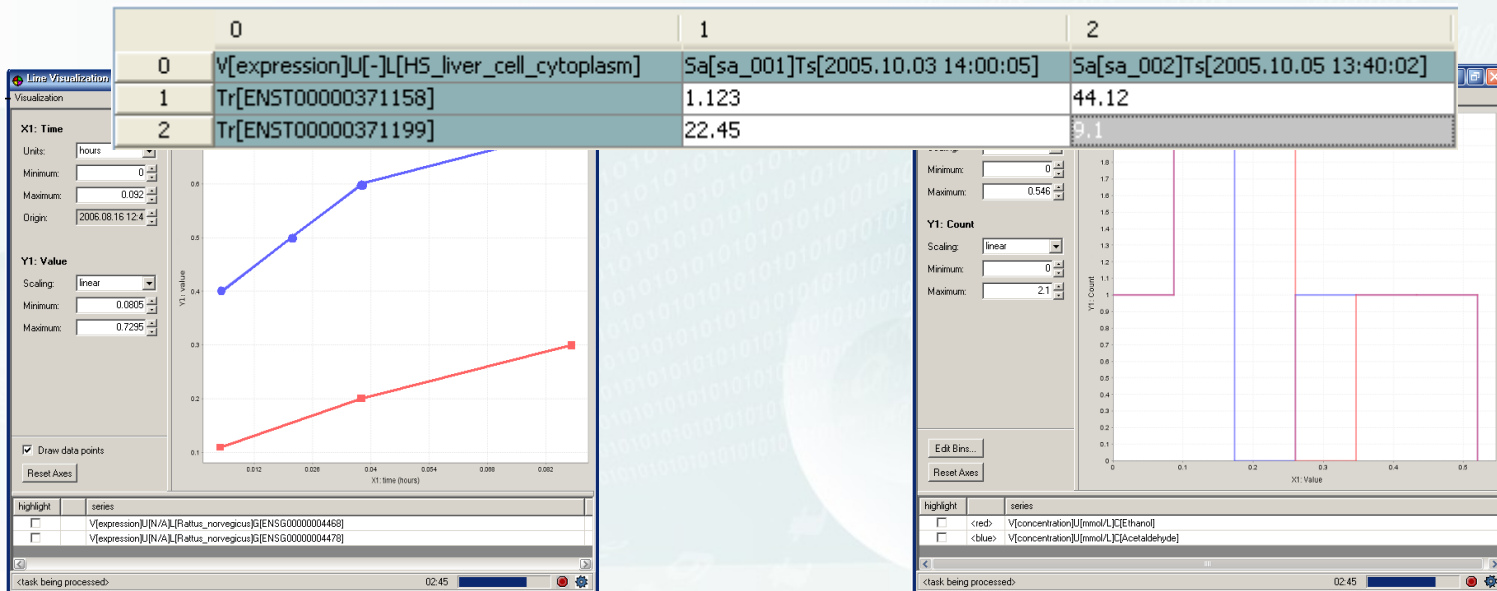


State data processing in workflows

- Basic Data set manipulation tools
 - Splitting and combining data sets
- Basic calculations
 - Tools for most common operations provided
- Data analysis tools
 - Clustering
 - Statistical testing
- Arbitrary calculations on data:
 - More advanced calculations can be performed using
 - R / Matlab programs
 - Perl/Python/... scripting

State data in workflows (cont'd)

- Basic viewers for viewing numerical data contents
 - Matrix view into data sets
 - Simple line plots



- For more complex viewing : State application

Part 3 – Where does state data come from?

Sources of state data

- Wet-lab measurements
- Publications / supplementary data
- Simulation
- External databases (via database population)

Wet-lab measurements

- Currently, specifically designed support for
 - Affymetrix (MASv5.0 + .CEL)
 - ABI, Illumina
 - Single shot images (LM & EM)
 - Image stack of single/multiple channels (LM & EM)
 - Mosaic microscope images (LM & EM)
 - LC-MS² (liquid chromatography tandem mass spectrometry) spectra in PKL format
 - Glycopeptide LC-MS² spectra in PKL, DTA or MGF format

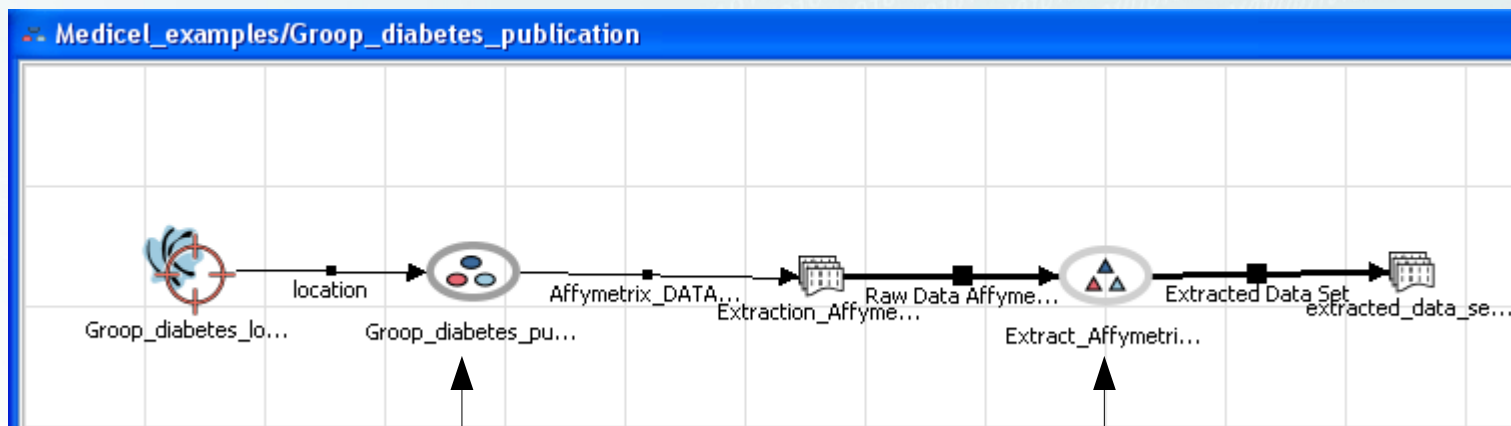
Example: Wet-lab measurement

- How to use Affymetrix .CEL files in Integrator?
- 5 steps
 - 1) Create a new wet-lab project
 - 2) Instantiate a predefined wet-lab template
 - 3) Import the raw (CEL) file(s) into Integrator
 - 4) Instantiate a predefined extraction workflow
 - 5) Execute the workflow

Publication supplementary data

Example:

M ootha et al. 2005: PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes



Minimalistic Affymetrix template,
in order to input the metadata

Minimalistic Affymetrix workflow,
in order to extract the raw data

Summary

Summary

- State data:
 - values,
 - for a quantity,
 - related to a component of a system,
 - at a given time
 - measurements
 - simulation results
- State data comes from
 - External DBs, wet-lab measurements, published supplementaries, simulation results, ...

Thank you!

- contact support@medicel.com
 - for all questions, problems etc.

<http://pilotapp1.medicel.com:6060/integrator/default.html>