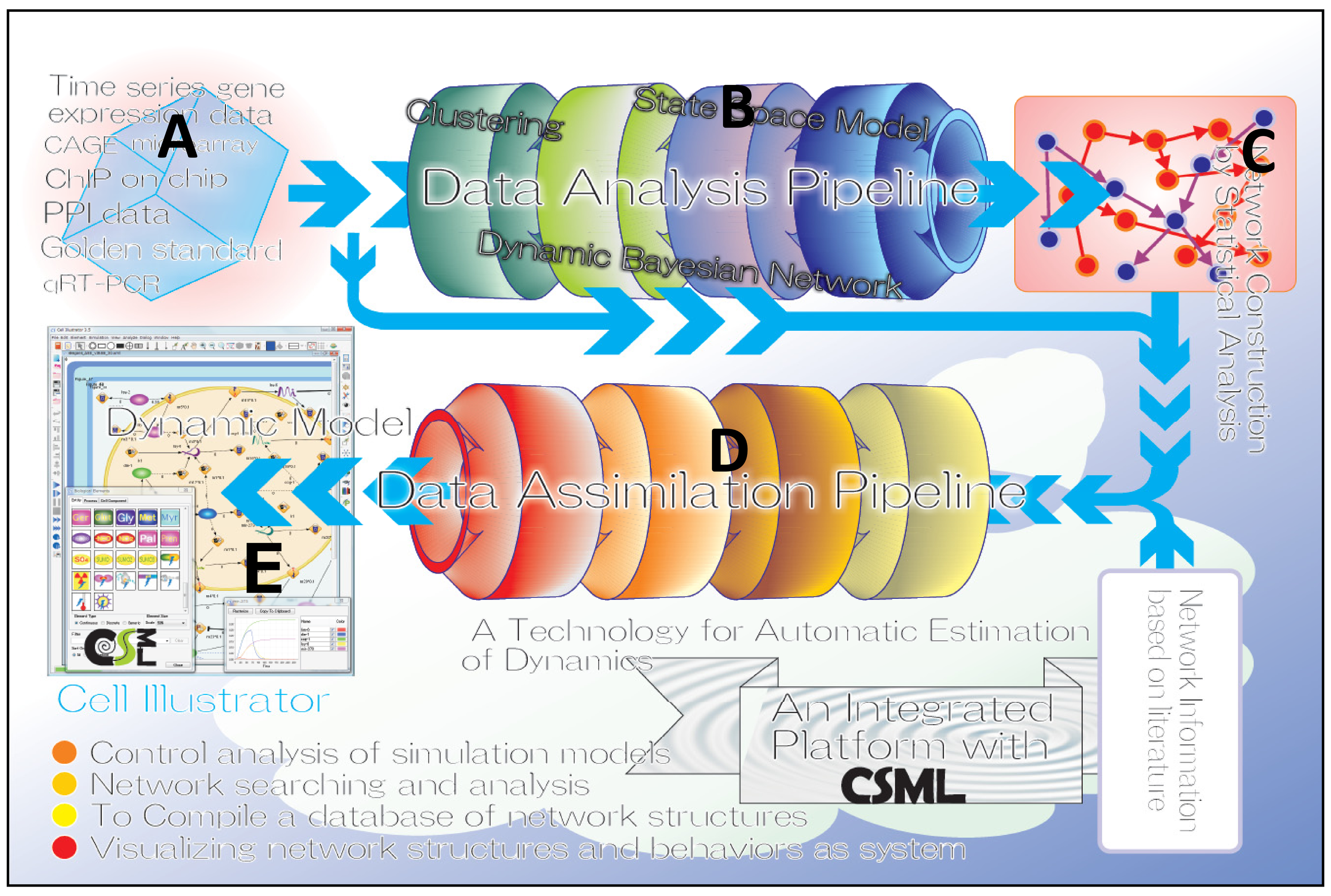
宮野　悟
東京大学医科学研究所

動的ネットワーク抽出のためのイン・シリコパイプラインの構築

Nagasaki M. , Imoto S., Fujita A., Saito A. , Ikeda E. , Shimamura T., Yamaguchi R., Ueno K., Miyano S.
Human Genome Center, Institute of Medical Science, University of Tokyo



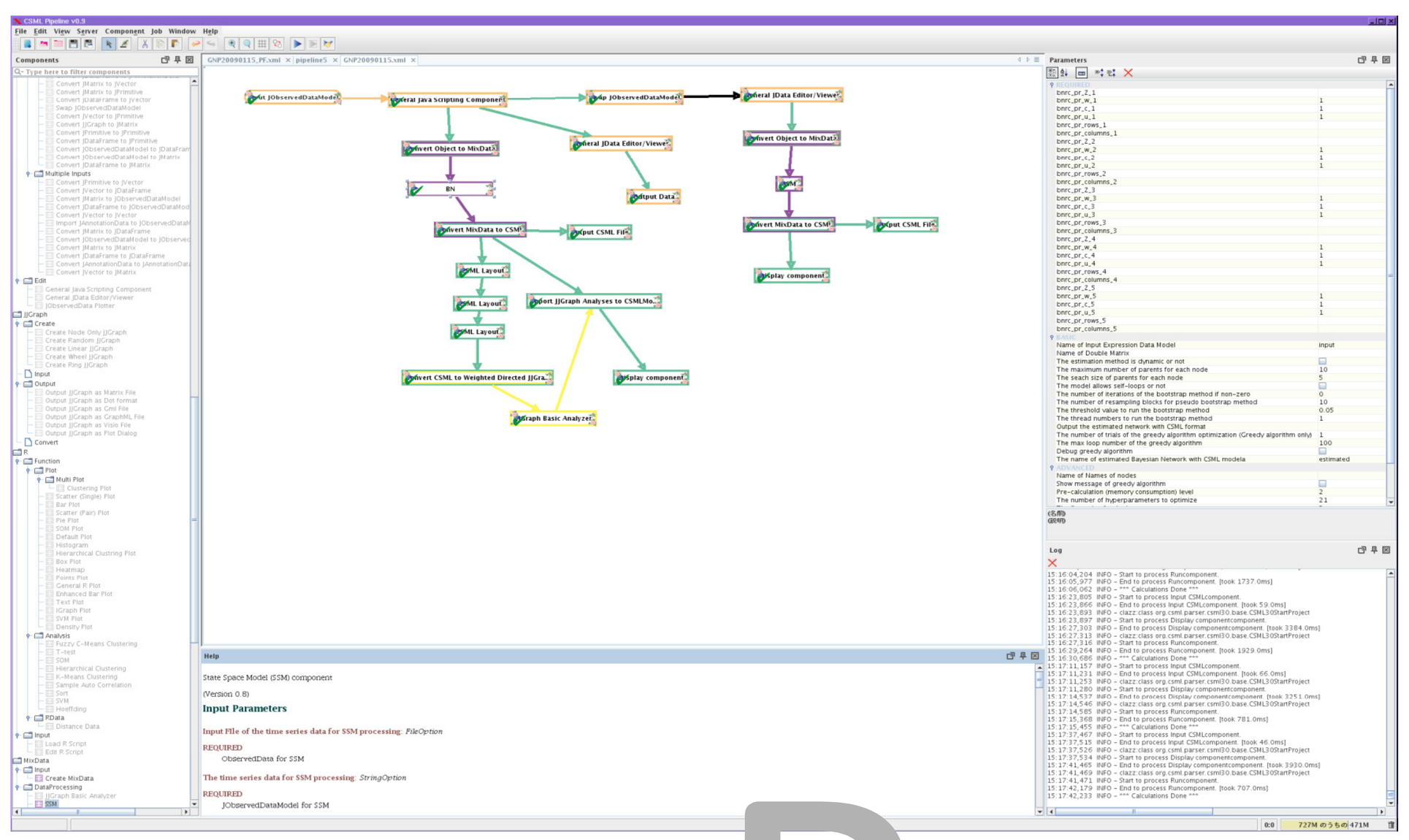
Abstract: Several technologies are currently used for gene expression profiling such as Quantitative Real Time RT-PCR (qRT-PCR), illumina microarray and Cap Analysis of Gene Expression (CAGE). CAGE is a recently developed method for constructing transcriptome maps and since its invention, it has been successfully applied to analyzing gene expressions in diverse biological studies. The principle of CAGE has been developed to address specific issues such as determination of transcriptional starting sites, the study of promoter regions and identification of new transcripts. Here, we present a Cell System Markup Language (CSML) computational pipeline to analyze high-throughput data derived from CAGE, microarray and qRT-PCR. This pipeline performs both standard statistical analysis and the most recently published methods of construction of regulatory networks and data assimilation.

CSML pipeline scheme. CSML pipeline receives as input, an EDF (Expression Data Format) file and performs several computational and statistical analysis including construction of regulatory networks and data assimilation. Results may be visualized using Cell Illustrator.

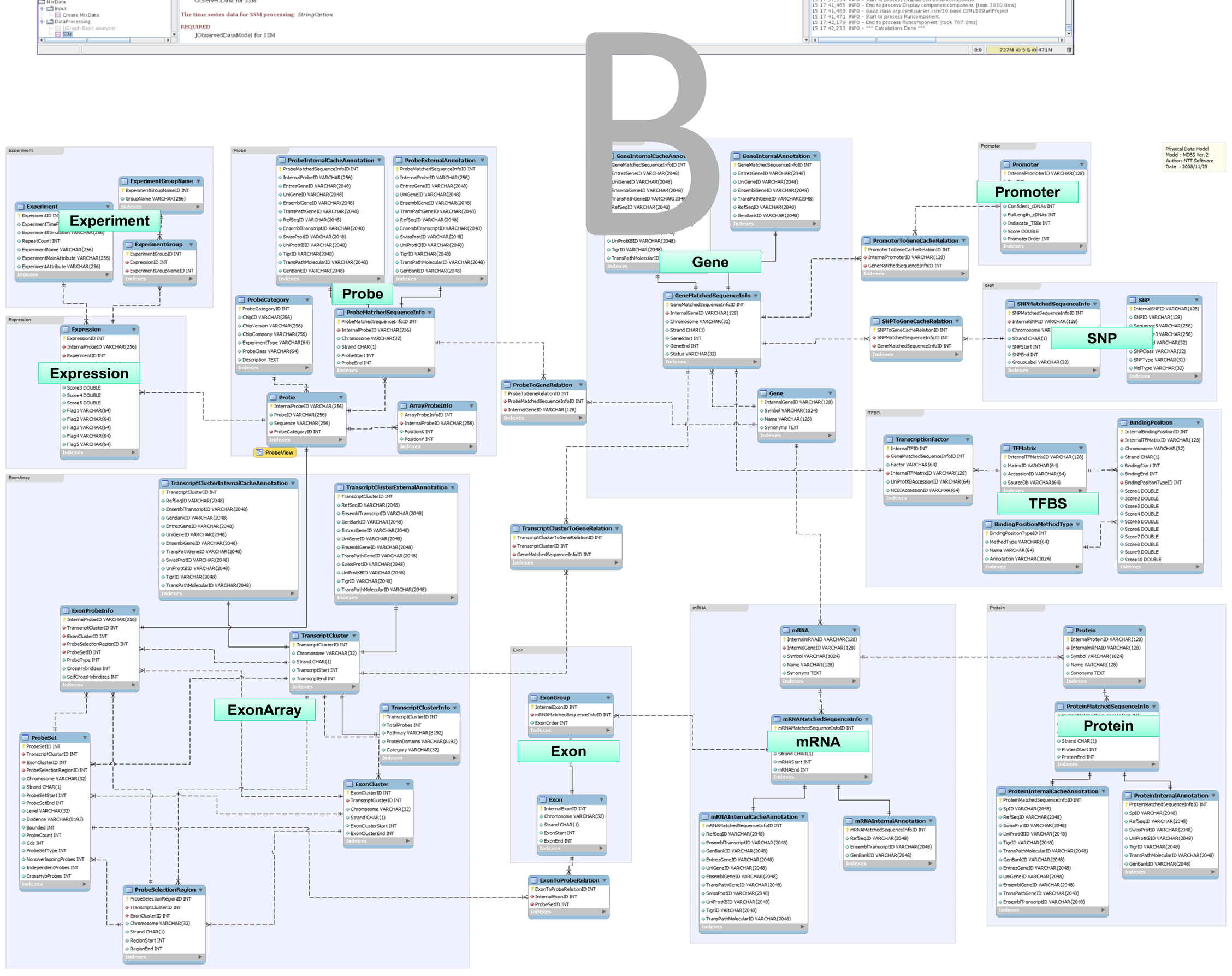
A

$encoding	UTF-8														
$version	1														
$primarykey	time														
$replicateID	repID														
$typeID	type														
$experimentType	microarray														
$chipID															
$chipVersion															
$chipCompany	illumina														
$description															
$date	12/2/2008														
@repID	1	1	1	1	1	1	1	1	1	1	1	1	1		1
@time	0	1	2	3	4	5	6	8	10	12	15	18	21		28
@type	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp	exp
6960451	54.50	33.31	21.00	33.93	41.85	22.63	60.35	40.71	60.06	87.05	130.68	131.71	111.81	94.13	133.06
2850504	2.82	2.76	3.63	2.74	2.86	2.94	3.59	3.76	3.07	3.89	3.68	2.88	2.58	3.89	3.23
7560397	3.49	3.33	2.89	2.93	2.63	3.18	4.79	5.68	5.50	3.54	7.78	4.17	3.10	3.52	2.85
5360747	3.33	2.42	2.50	2.69	3.07	2.38	3.62	2.64	2.15	2.57	2.78	3.16	2.93	3.50	2.79
2600731	195.14	594.52	810.50	548.09	351.30	393.13	361.60	492.52	412.80	375.94	252.54	319.30	248.54	281.30	170.64
3990110	5.82	6.61	5.12	4.77	6.92	4.49	4.29	4.19	3.82	4.05	3.17	4.94	4.82	4.17	5.89
2120309	1.78	1.85	1.86	2.40	2.11	2.21	2.22	2.79	2.10	2.50	1.27	2.61	1.93	2.83	2.27
6420474	55.43	33.31	39.72	45.25	52.73	25.88	45.33	26.68	26.29	44.56	57.49	45.68	40.73	48.42	58.12
670095	3.05	2.59	2.96	2.51	4.75	3.59	2.17	2.46	2.06	2.13	2.75	2.08	2.33	2.20	2.28
130367	6.74	4.97	2.98	5.68	6.85	3.38	4.63	4.38	4.05	4.85	3.66	4.65	3.82	5.47	5.94
.															
.															
.															

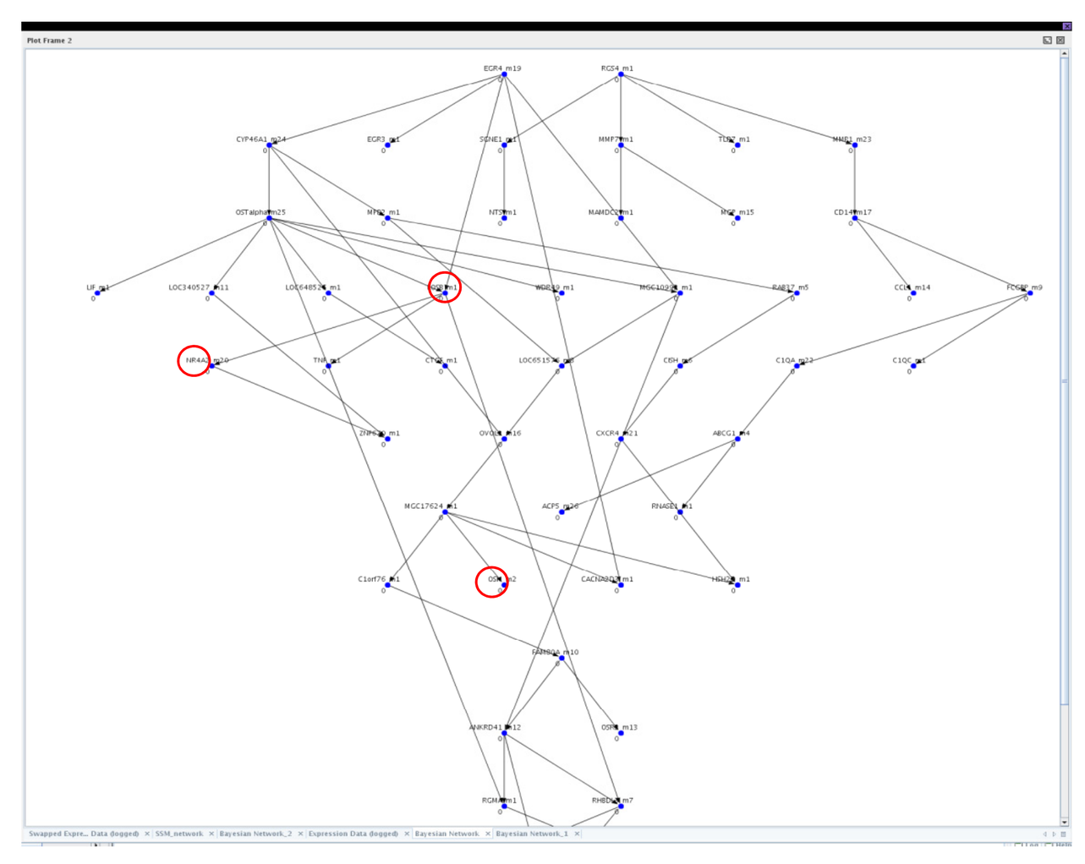
Input file. The input is an EDF (Expression Data Format) file which contains information about the technology used in the experiment in the header (initialized by the "$" symbol), information about the type of experiment in the attributes (initialized by the "@" symbol) and finally the expression data. If a line is initialized with the symbol "#", it comments out the respective line.

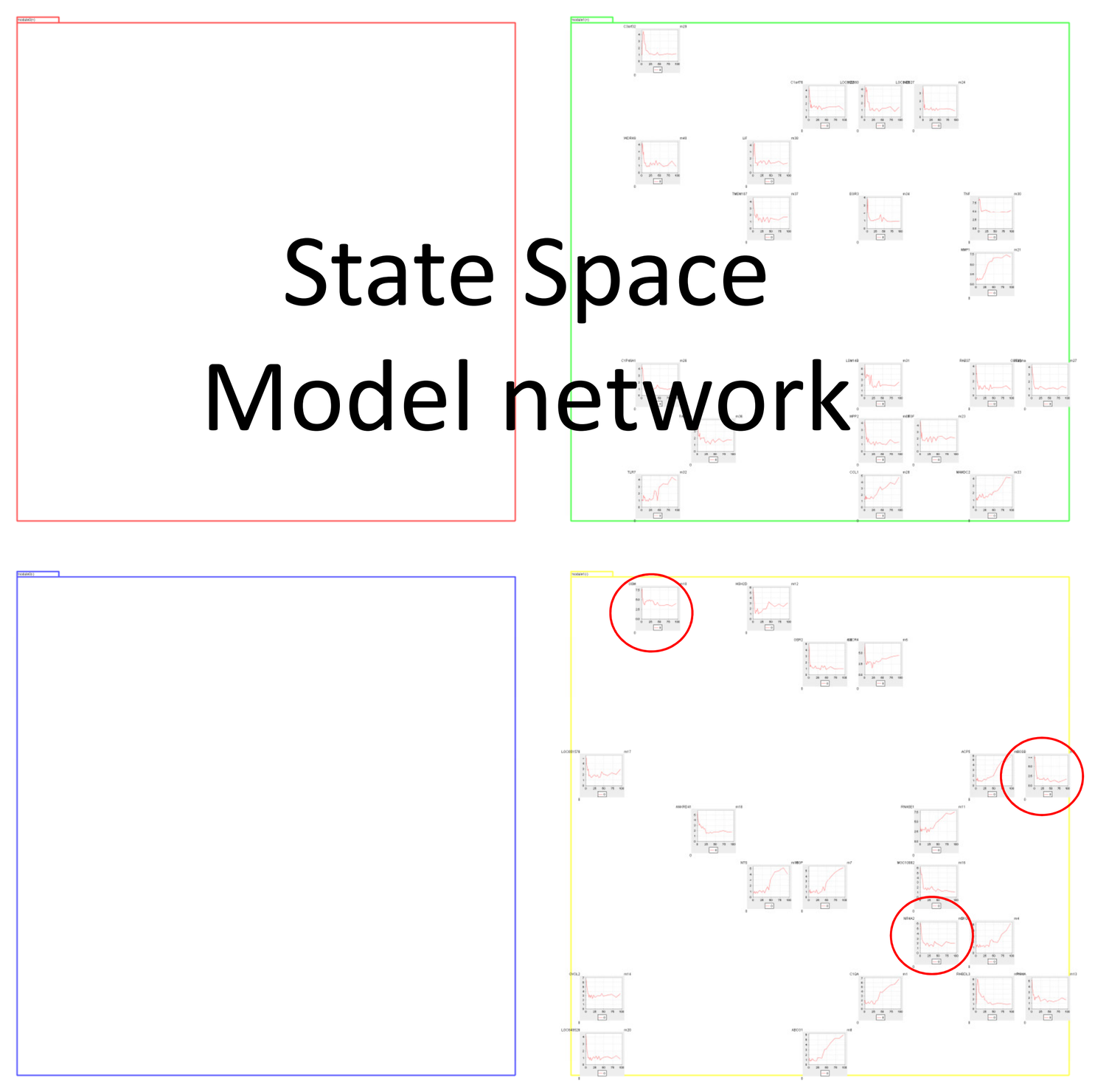


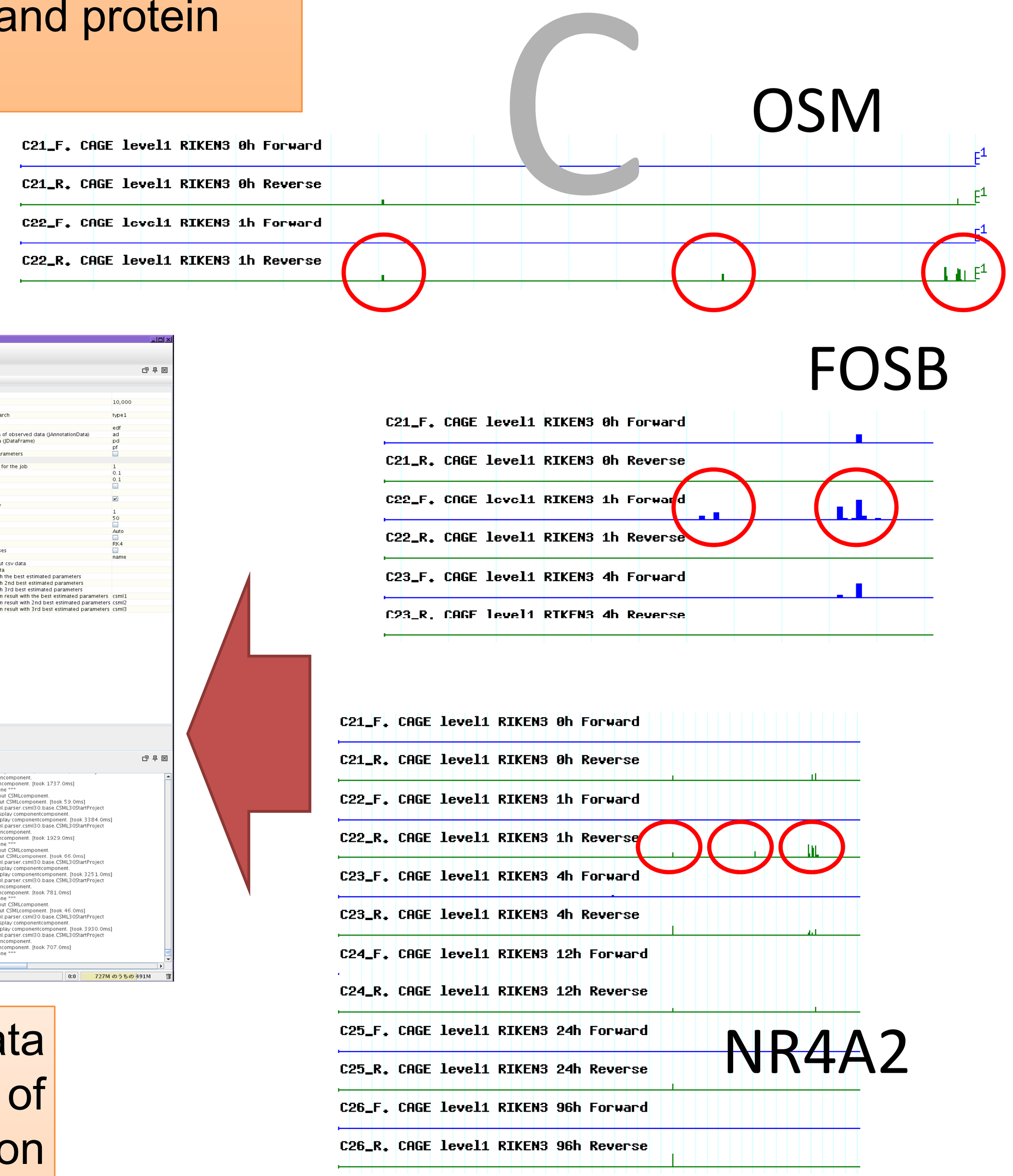
CSML pipeline graphical interface. The user constructs himself the pipeline, just dragging and dropping the components.

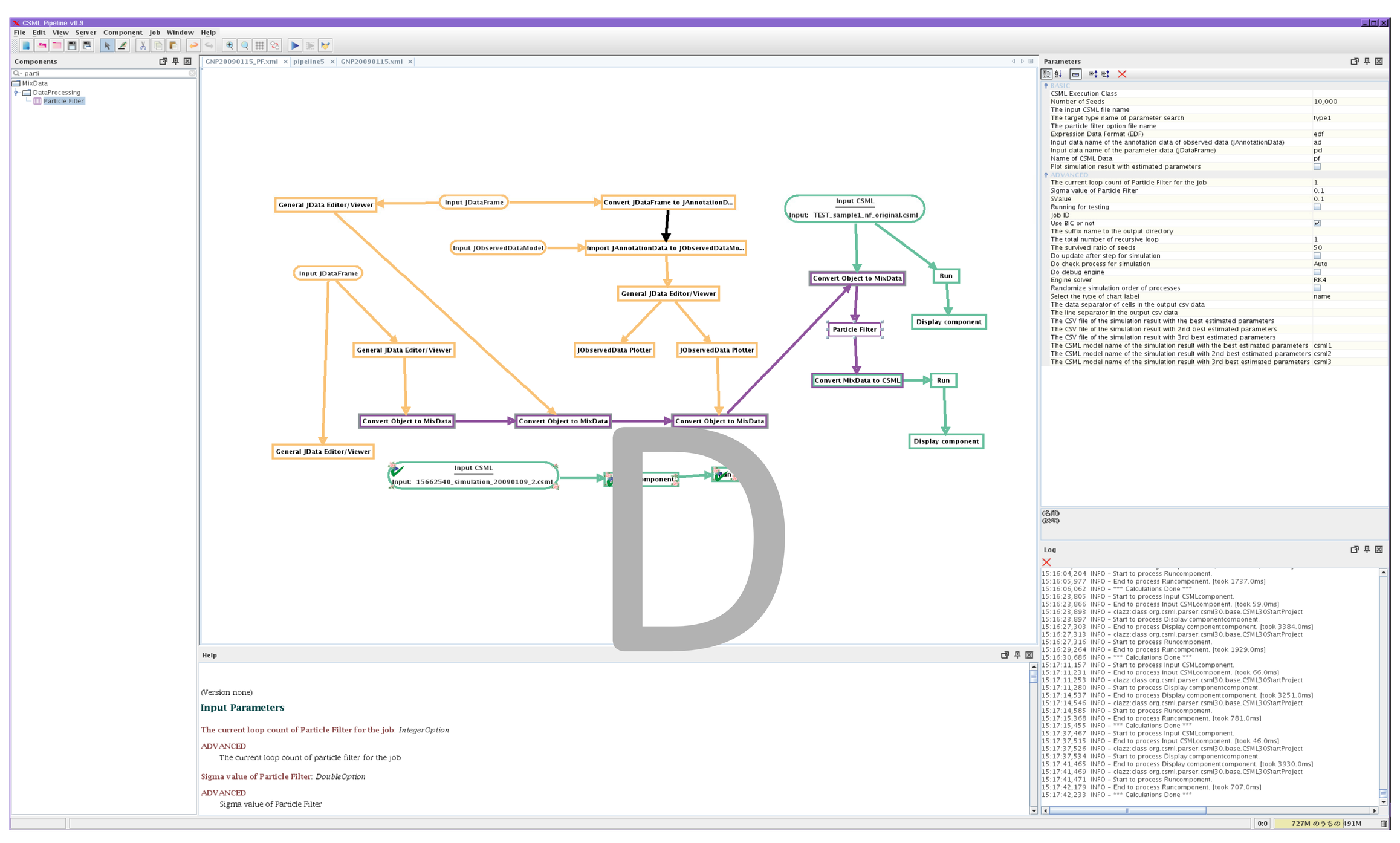


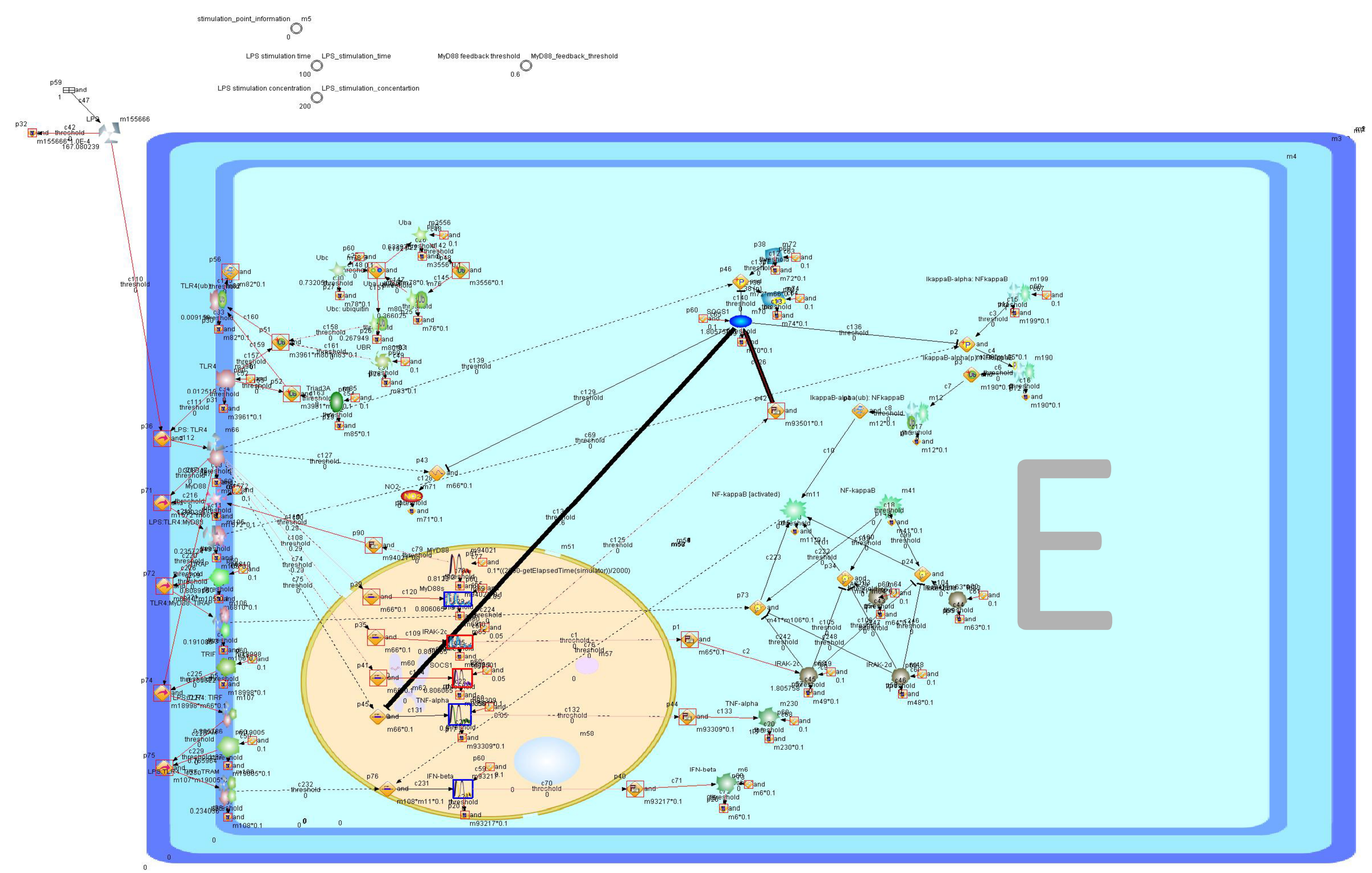
Database scheme. In the actual version (version 2), this database contains information related to experiment, probe, gene, promoter region, expression level, transcription factor binding sites, SNP, exon, proteins, mRNA and ExonArray. The database is easily extendable , and future versions aim to include information about gene ontology and protein arrays.

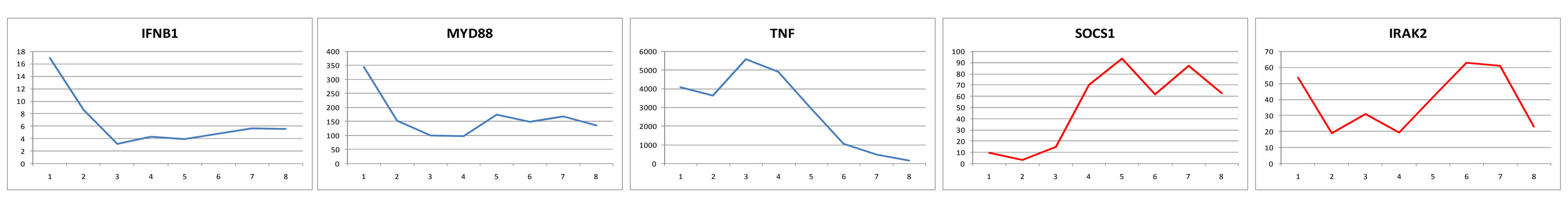


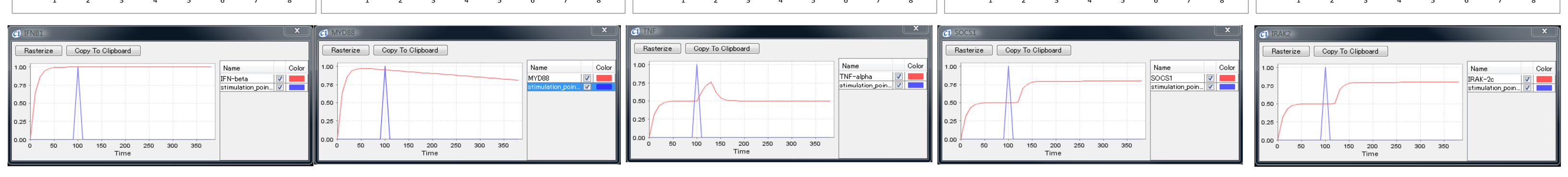












Dynamic model. The output is visualized using Cell Illustrator. Additional simulations and analysis will be performed in this platform. The network is the signal transduction and gene regulatory network model with the LPS stimulation to macrophage.

Data Assimilation pipeline. Data Assimilation technology (combination of literature data, a priori biological information and mathematical model) is used to improve the biological pathway model to be more reasonable pathway model by using the observed data, e.g. mRNA expression data.

CAGE expression data. CAGE technology permits to visualize a network of transcripts instead of a network of genes.

Supported by: Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa