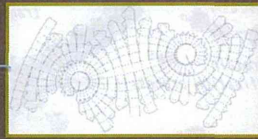
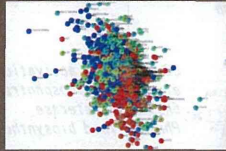


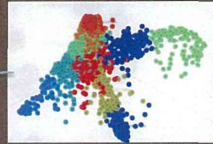
## Processing data on AGCT

1. 時系列データ前処理: 線形回帰/ウェーブレット変換
2. 遺伝子間の類似度マトリックス
3. 低次元に落とすためにSpectral clusteringを行う。通常の主成分分析も行う。
4. 発見的なClustering法を使って構造上でデータの分割を行う。
5. 結果のinteractive visualizationやscenario 記録を行う。

PCA :  $M \times N$  matrix



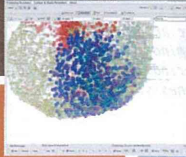
Spectral clustering:  
 $M \times M$  matrix



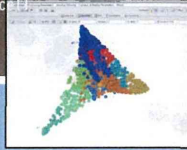
orthogonal matrix to compute  
one dimension per cluster/gene

## Examples of different network topologies

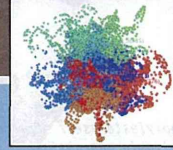
Mouse Stem cell



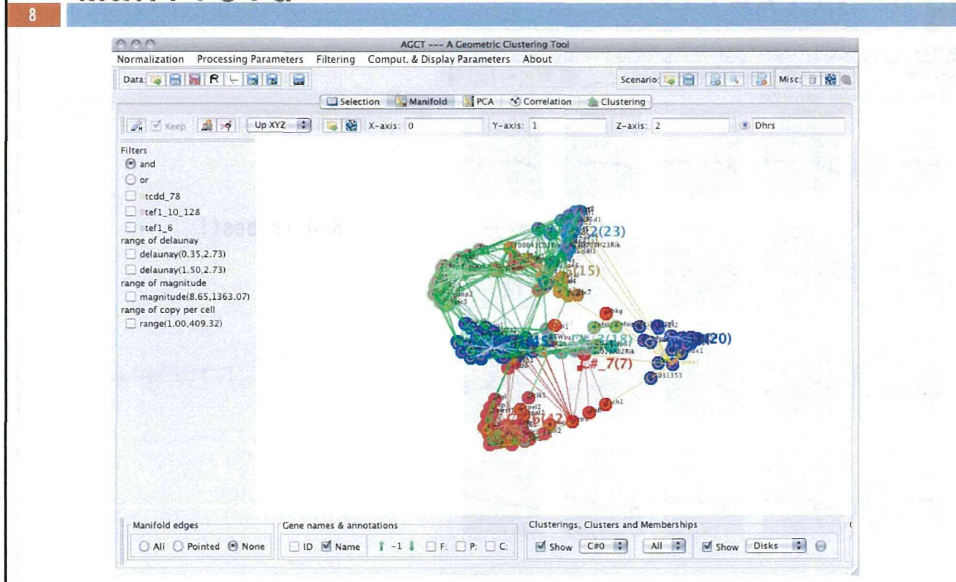
TCDD affected mouse liver cell



Influenza affected mouse

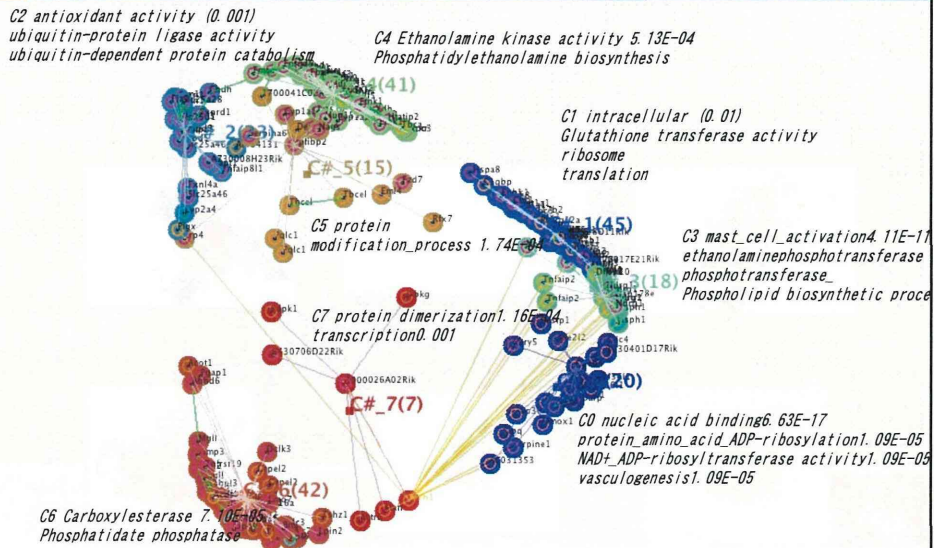


## AGCT result: visualization on Manifold



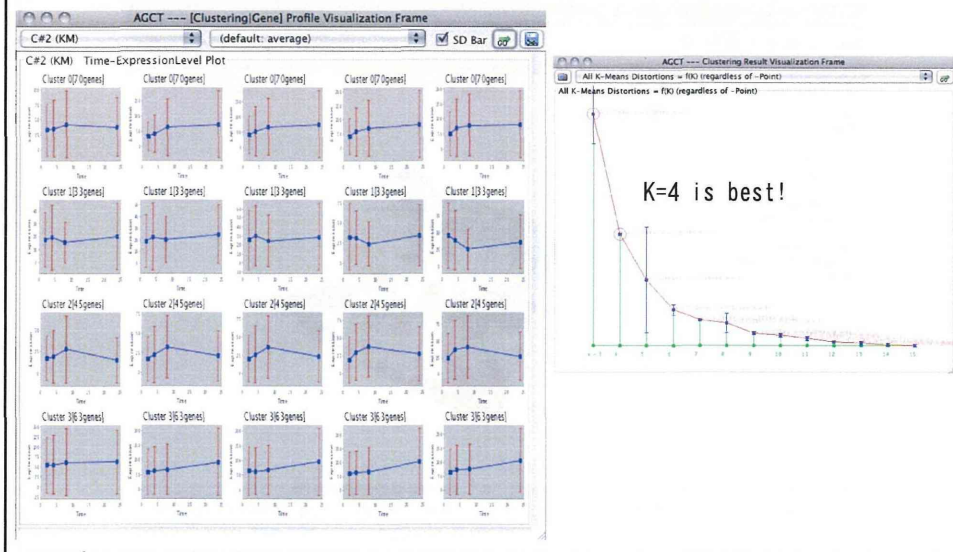
# Computing gene ontology for Clusters

9



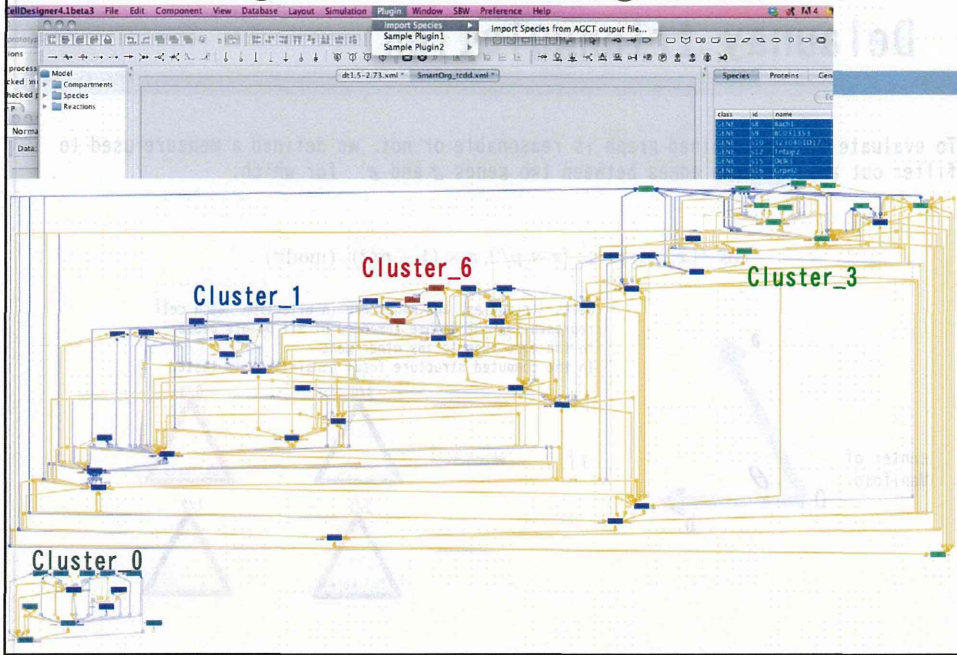
# Selection of cluster number and visualization

10

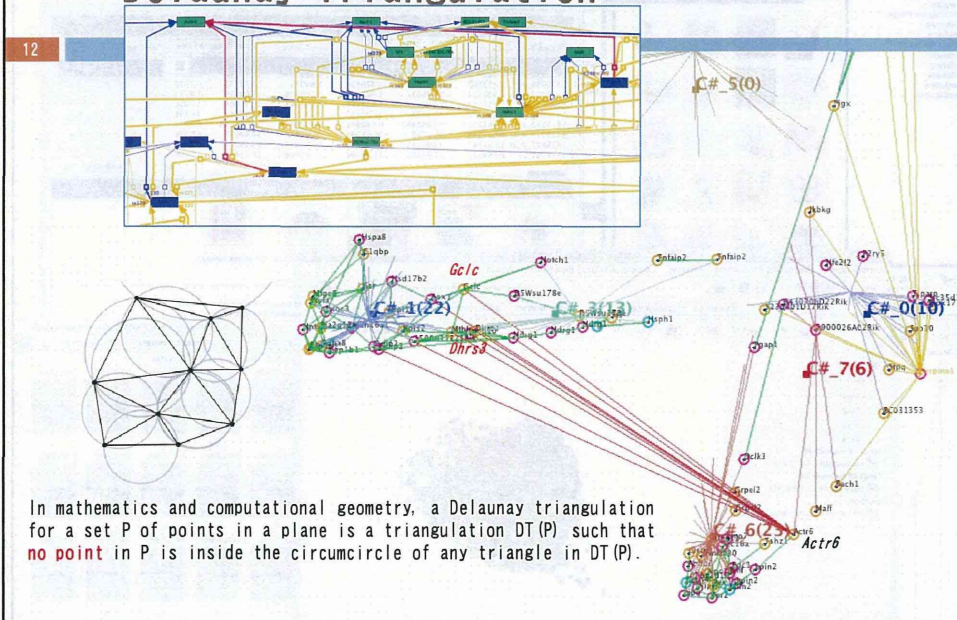




# AGCT Plugin to CellDesigner



## AGCT clusters in CellDesigner by Delaunay Triangulation

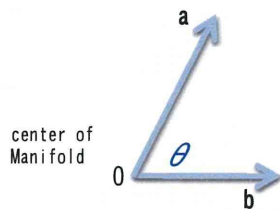


# Delaunay triangulation

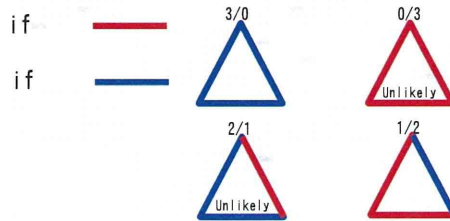
13

To evaluate if the obtained graph is reasonable or not, we defined a measure used to filter out any Delaunay edges between two genes  $g$  and  $g'$  for which:

$$\cos^{-1}(\mathbf{x}_g, \mathbf{x}_{g'}) \in [\pi \times p/2, \pi \times (1 - p/2)] \pmod{\pi} .$$



Cells (triangles) and genes are in bijection, each cell representing the volume composed of all points closer to the gene than to any other gene. In the computed structure local consistency is tested.




# Connection to Garuda

The screenshot displays the Garuda software interface, which includes a sidebar with categories like 'A Starter Kit', 'Analytics', and 'Clustering'. The main window shows a 'Garuda Trace' window with a table of data. Below this, there is a 'Garuda Discovery Engine' section. In the foreground, the 'AGCT - A Geometric Clustering Tool' window is open, showing a 3D visualization of a gene network with nodes and edges. The interface also includes various filters and parameters for data analysis.

File Contents	1.999515	2.972562	4.927777
1415046_at	1.248562	2.991156	5.470451
1415004_s_at	1.901521	2.945337	1.908072
1416234_at	1.755584	1.741552	1.619864
1416005_at	18.64732	12.8408	35.2511
1419152_at	0.4612821	0.7382492	0.5889461
1415916_a_at	72.22558	46.2718	55.55189
1418213_s_at	11.8174	9.957685	12.26441
1434565_at	20.89522	15.73216	17.46994



15




# SHOE


- Sequence Homology in Higher Eukaryotes

# SHOE Interface

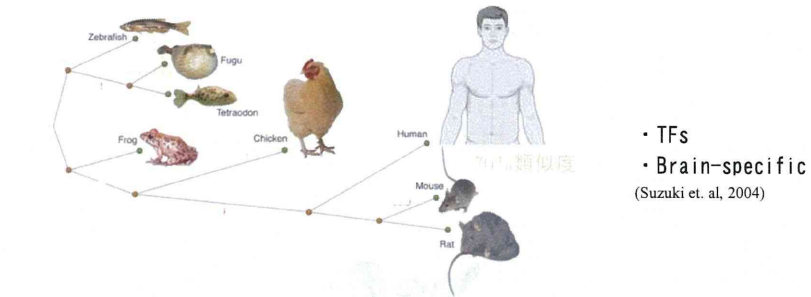
16



Sony CSL Server



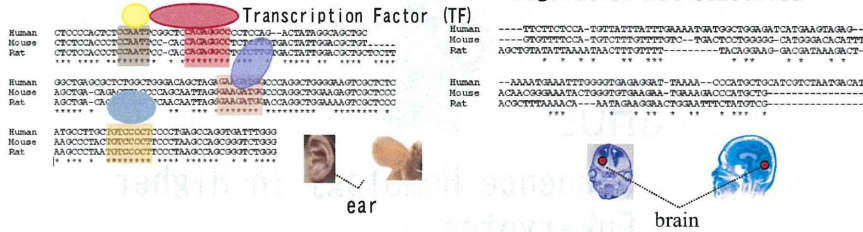
## Phylogenetic footprinting finds evidence of functionality



- TFs
  - Brain-specific
- (Suzuki et. al, 2004)

Regulation conserved!

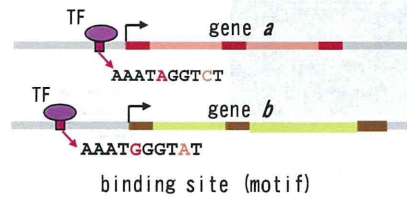
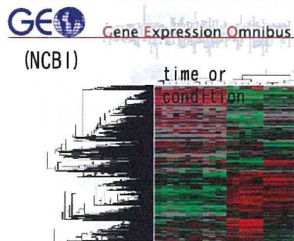
Regulation not conserved



17

## Co-regulated genes

Similar way of expression suggests **regulation** by the same Transcription Factor (TF)



Motif discovery tools

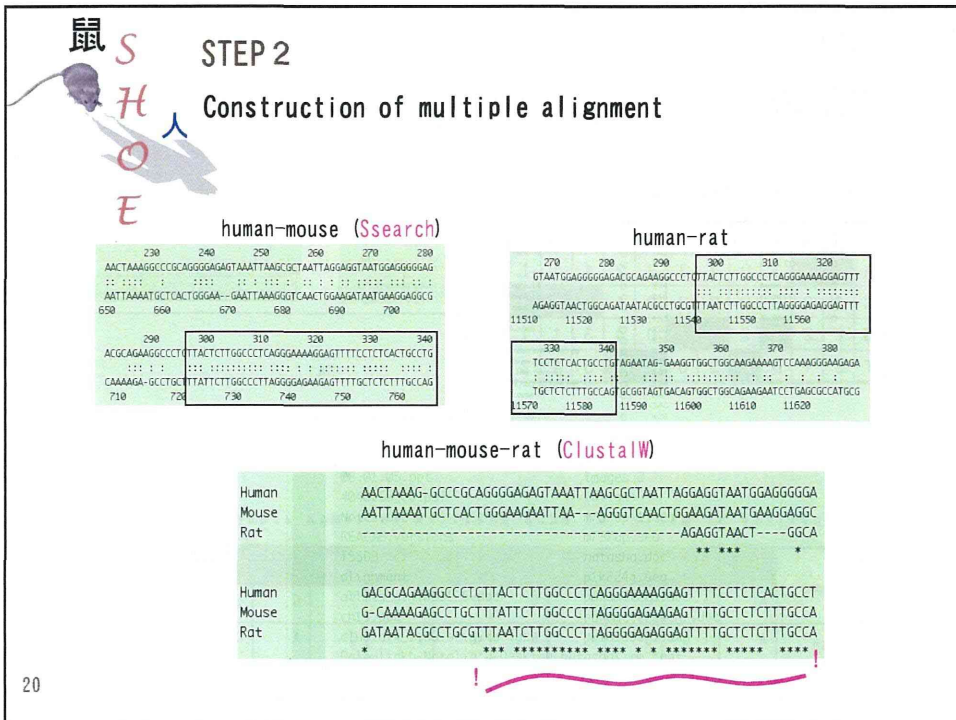
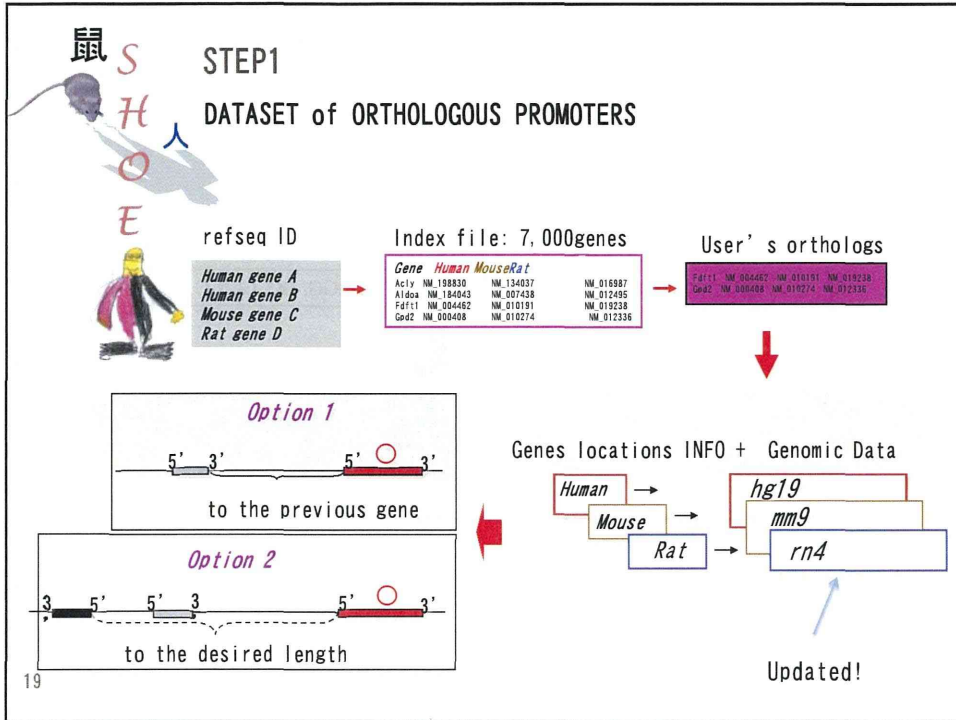
&

Comparative genomics tools

MEME (Bailey *et al.*)  
 Consensus (Stormo *et al.*)  
 Gibbs sampler (Lawrence *et al.*)  
 Yebis (Yada *et al.*)....

MONKEY (Moses *et al.*)  
 FootPrinter (Tompa *et al.*)  
 PhyMe (Sinha *et al.*)  
 PhyloGibbs (Siddharthan *et al.*) ....





STEP 3

How good is the alignment?

Pattern frequencies tables for "good" and "random" alignments  
74 patterns

A, T, G, C, - × AA, AT, AG, AC, A-, TT, TG, TC, T-, GG, GC, G-, CC, C-, --

835 orthologous alignments  
(238,800bp)

1260 random alignments  
(239,600bp)



21

$c$  - probability of pattern in each column in *good alignment* and *random alignment* tables,  
 $m$  - motif length.

**BIODATABASES**  
BIOLOGICAL DATABASES

**TRANSFAC**

STEP 4

Calculation of PSSM score ( $PM_{score}$ )

498 human-mouse rat matrices

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪
A	4	5	3	0	4	3	3	2	1	1	1
C	1	2	0	0	0	0	0	1	3	4	6
G	2	2	7	2	3	7	0	4	3	1	1
T	3	1	0	8	3	0	7	3	3	4	2

10

where  $pseudocount = 1$   
 $m$  is a motif length

Probability of motif at each position in human sequence

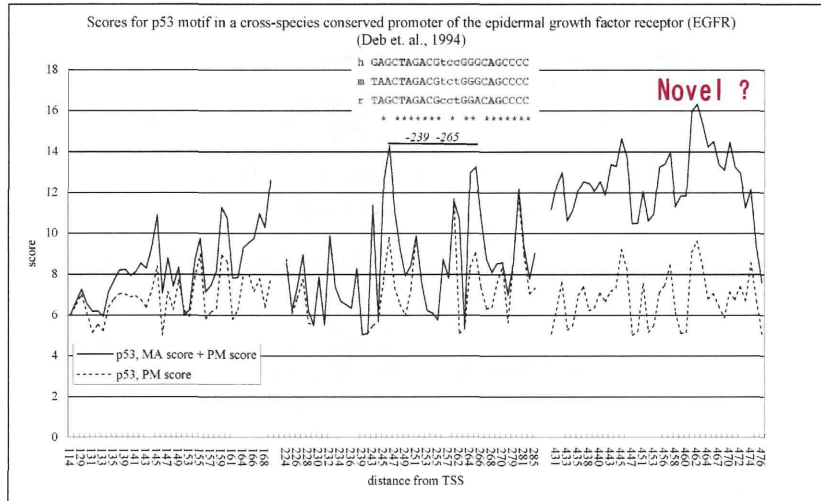
human	-0.48	-0.39	-0.27	-0.22	-0.57	-0.27	-0.27	-0.57	-0.57	-0.47	-0.33	-4.46
A	A	G	T	G	G	T	T	C	C	C		
mouse	A	T	G	T	G	G	T	C	A	C	C	
rat	A	T	C	T	G	G	T	A	A	C	C	

22



# Finding motif candidates

Multiple alignment score (MA) + Transfac PSSM score (PM)



23

# SHOE development

BrowserTest

SONY Sony CSL **SHOE** Sequence Homology in higher Eukaryotes Job Input Queue List

No	Start Date	End Date	Gene List	Repeat Masker	Upstream Length	Mode	Downstream Length	Scoring
90	2013/05/29 15:58:54	2013/05/29 16:09:46	NM_001002239 NM_001013785	Checked	5000	mode2	500	transfac32

90 PROMOTER (14 genes)  
 ① Genes RPL17  
 ~7,000個

ALIGNMENT (11 genes)  
 ② Genes orthologous

File Name: NM\_001002239\_5000(47023276-47023862)\_1aln.fasta.result (224,797 lines) Location to TSS

PSSM	TF	Location to TSS	Str	MA Score	PSSM Score	MA Score + PSSM Score (Threshold)
M00193	NF-1	211 ~ 228	+1	9.30	-9.64	19.95
M00991	CDX	210 ~ 227	+1	9.04	-9.62	18.67
M00531	NERF1a	214 ~ 231	+1	8.89	-9.92	18.82
M01023	HSF1	212 ~ 228	+1	8.73	-9.68	18.42
M00528	PPAR	210 ~ 226	+1	8.69	-9.03	17.73
M00528	PPAR	211 ~ 227	+1	8.69	-9.90	18.60
M00526	GCFN	217 ~ 234	+1	8.63	-8.07	16.71

# Alignment viewing

BrowserTest

SONY Sony CSL **SHOE** Sequence *HQ*mology in higher Eukaryotes Job Input Queue List

No	Start Date	End Date	Gene List	Repeat Masker	Upstream Length	Mode	Downstream Length	Scoring
90	2013/05/29 15:58:54	2013/05/29 16:09:46	NM_001002239 NM_001013785 NM_001026214	Checked	5000	mode2	500	transfac:32

File Name: NM\_001002239\_5000[47023276-47023862]\_1aln.fasta.result (224,797 lines.) Location to TSS

LIMIT 50

PSSM	TF	Location to TSS	Str	MA Score	PSSM Score	Threshold
M00193	NF-1	211 ~ 228	+1	9.30	-9.64	18.95
M00991	CDX	210 ~ 227	+1	9.04	-9.62	18.67
M00531	NERF1a	214 ~ 231	+1	8.89	-9.92	18.82
M01072	WCE1	717 ~ 738	-1	8.72	-9.60	18.47

# Motif result viewing

BrowserTest

SONY Sony CSL **SHOE** Sequence *HQ*mology in higher Eukaryotes Job Input Queue List

No	Start Date	End Date	Gene List	Repeat Masker	Upstream Length	Mode	Downstream Length	Scoring
97	2013/06/04 11:48:05	2013/06/04 11:53:53	NM_007804 NM_011400	Checked	5000	mode2	500	transfac:32

File Name: NM\_011400\_5000[43425277-43426579]\_1aln Location to TSS

TSS: 43429347

Location to TSS (Human)

CLUSTAL W (1.83) multiple sequence alignment

File Name: NM\_011400\_5000[43425277-43426579]\_1aln.fasta.result (510,760 lines.) Location to TSS

LIMIT 50

PSSM	TF	Location to TSS	Str	MA Score	PSSM Score	Threshold
M001	GR	292 ~ 310	-1	9.26	-9.40	18.67
M002	11:MatG	236 ~ 257	+1	9.19	-9.96	19.16
M001	Oct-1	50 ~ 68	+1	8.98	-8.97	17.96
M005	NERF1a	156 ~ 173	+1	8.92	-9.42	18.35
M00005	AP-4	583 ~ 600	-1	8.84	-9.42	18.27
M00991	CDX	584 ~ 601	-1	8.84	-9.02	17.86
M01007	SRF	281 ~ 299	-1	8.83	-9.40	18.24



# Selection of results 結果の絞り込み

1 2

27

SONY Sony CSL SHOE Sequence Homology in higher Eukaryotes

Job Input Queue List

No	Start Date	End Date	Gene List	Repeat Masker	Upstream Length	Mode	Downstream Length	Scoring	
97	2013/06/04 11:48:05	2013/06/04 11:53:53	NM_0078 NM_0114	10 100 300 500 1000 5000	Checked	5000	mode2	500	transfac32

File No: 100\_5000(43425317\_43426579)\_1aln  
File No: 100\_5000(43425317\_43426579)\_1aln.fasta.result (510,760 lines)

LIMIT: 10000  
PSS: 50000 no limit  
Location: to TSS

RESULT: 32 lines

MA Score ↑ PSM Score

Gene	TF	Position	Str.	MA Score	PSM Score	MA Score = absolute(PSM Score)
M00418	TCIF	591 ~ 601	+1	5.27	-2.84	8.12
M00184	MyoD	454 ~ 464	+1	5.05	-2.96	8.02
M00037	NFE2	241 ~ 251	-1	4.97	-2.90	7.88
M00983	MAF	241 ~ 251	-1	4.97	-2.37	7.35
M00469	AP-2alpha	1148 ~ 1156	+1	4.91	-2.81	7.73
M00470	AP-2gamma	1148 ~ 1156	+1	4.91	-2.84	7.76
M00174	AP-1	240 ~ 250	+1	4.67	-2.82	7.50
M00649	MAZ	514 ~ 521	+1	4.64	-2.81	7.45
M00342	Oct-1	55 ~ 64	+1	4.58	-2.71	7.30
M00712	myogenin	210 ~ 217	-1	4.57	-2.77	7.35
M00931	Sp-1	351 ~ 360	-1	4.53	-2.69	7.24
M00975	RFX	292 ~ 300	-1	4.48	-2.68	7.18
M00184	MyoD	36 ~ 45	-1	4.46	-2.79	7.25
M01008	Ebox	457 ~ 465	+1	4.42	-2.30	6.73
M01008	Ebox	1227 ~ 1235	-1	4.38	-2.24	6.63
M00332	Whn	274 ~ 284	+1	4.36	-2.86	7.22
M00933	Sp-1	205 ~ 214	+1	4.33	-2.75	7.09
M00217	USF	456 ~ 463	+1	4.28	-2.92	7.21
M00973	E2A	457 ~ 464	+1	4.26	-1.78	6.05
M00199	AP-1	319 ~ 327	-1	4.25	-2.61	6.88
M00712	myogenin	464 ~ 471	-1	4.19	-2.69	6.89
M00470	AP-2gamma	684 ~ 692	+1	4.16	-2.78	6.95
M00649	MAZ	465 ~ 472	-1	4.14	-2.59	6.75

all res SAVE

# Top scoring TF motifs with Pareto Front

パレート面を用いた多目的問題最適化

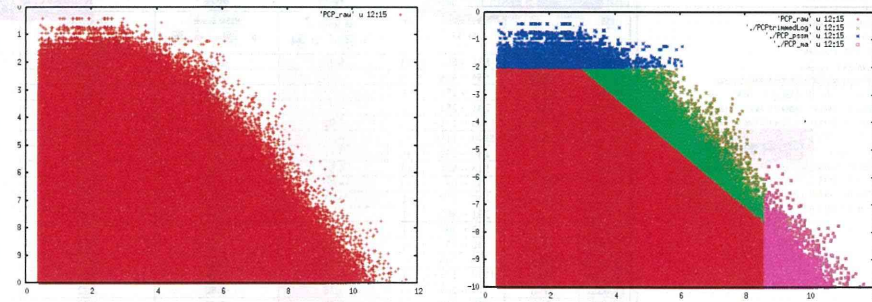
28

Mitsubishi Regional Jetおよび Bombardier CSeriesを支える次世代型エンジンの設計などに用いられ、工学的な応用が盛ん

Top scoring cross-region of two distributions: multiple alignment score and position specific matrix score

29

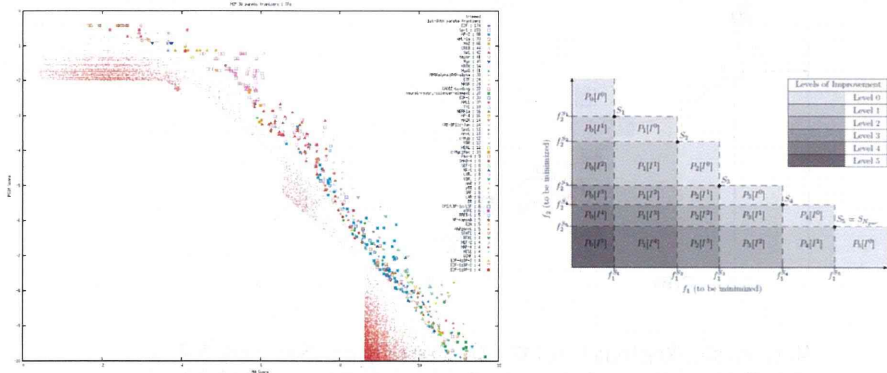
パレート面を用いた多目的問題最適化



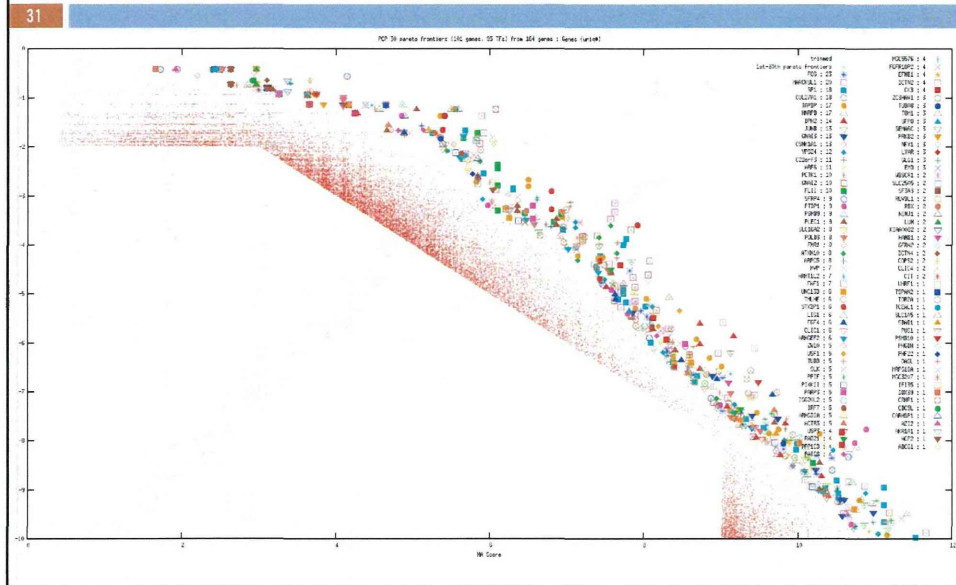
Top scoring cross-region of two distributions: multiple alignment score and position specific matrix score

30

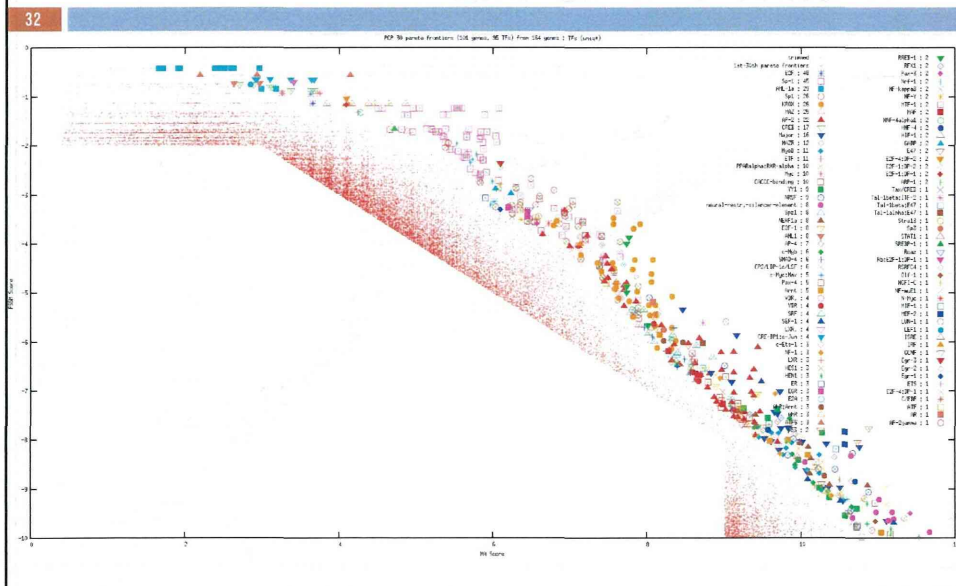
パレート面を用いた多目的問題最適化



## PCP network: Genes with unique number of top scoring TFs in each promoter



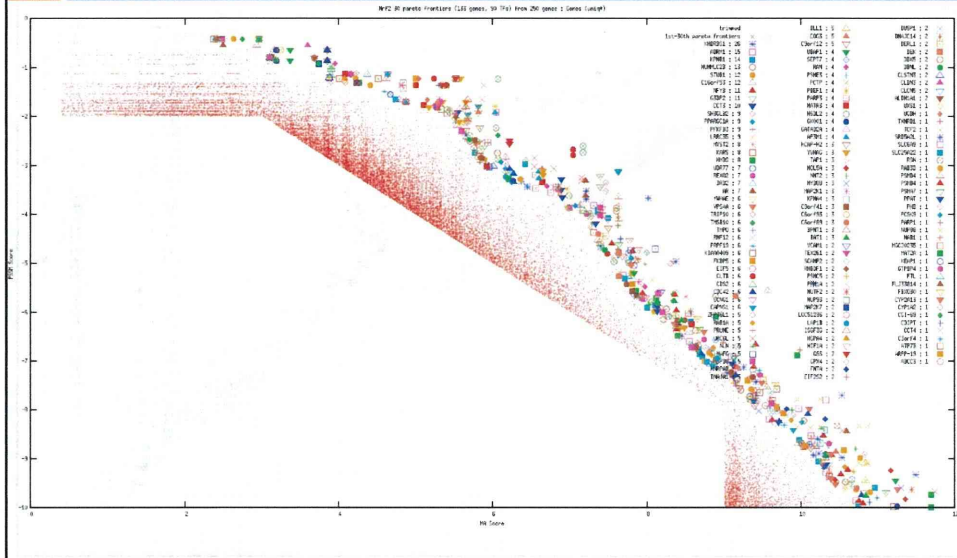
## PCPnetwork: TFs with a unique number of genes sharing them





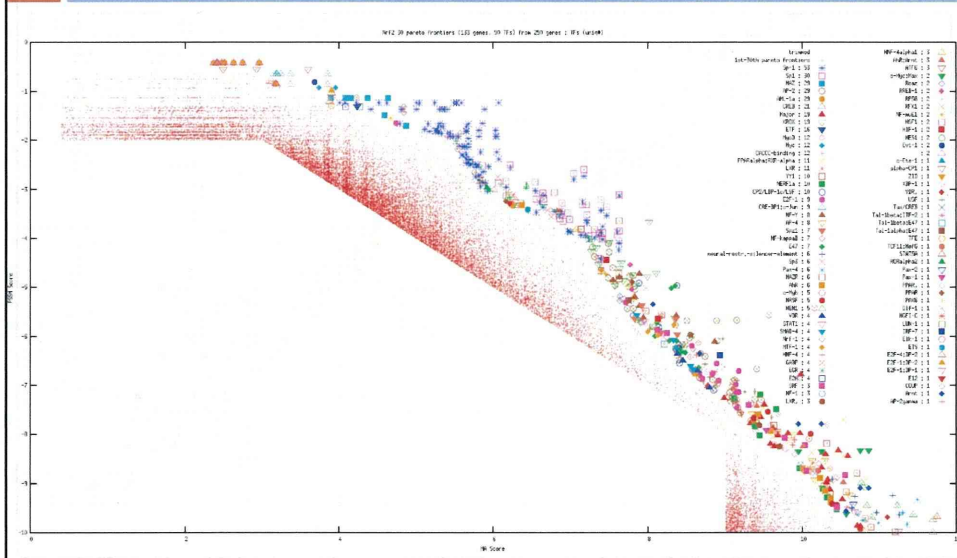
## Nrf2 network: Genes with unique number of top scoring TFs in each promoter

33



## Nrf2 network: TFs with a unique number of genes sharing them

34



## Co-authors

35

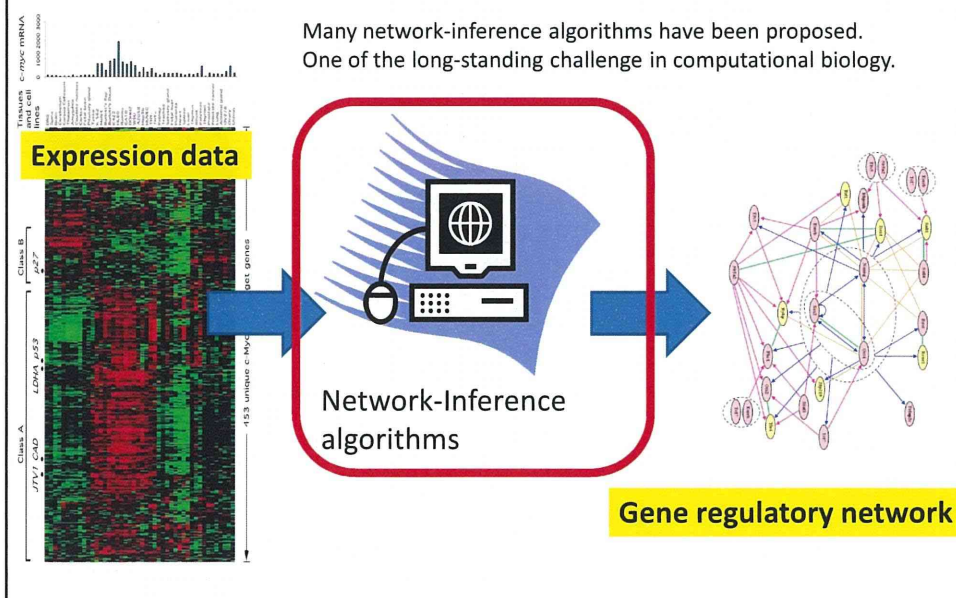
- Hiroaki Kitano (Sony CSL)
- Frank Nielsen (Sony CSL)
- Richard Nock (University of Martinique)
- Keigo Oka (University of Tokyo)
- Koudai Takata (Tokyo Institute of Technology)
- Tomohiro Masagaki (Sony CSL)

# Systems Toxicology

## Inference of gene regulatory networks from large-scale experimental data

Takeshi Hase, Samik Ghosh, and Hiroaki Kitano  
The Systems Biology Institute

## Inference of gene regulatory network





## Application of Network Reconstruction Technique to Percellome

- **Possible areas**
  - How network structure changes over time in a tissue for a specific dose of the perturbation?
  - Build tissue specific gene networks
  - How the gene interaction network changes over different dosage of the compound?

## Inference of gene regulatory network expression dataset under pentachlorophenol

## PCP (pentachlorophenol) dataset

- Gene expression data with PCP from mouse liver
- 4 time-points (2, 4, 8, and 24 hours) and 4 dosages (0, 10, 30, and 100 mg/kg PCP). Data of 2hour and 0 mg/kg Three replicates for each condition. 2 hours vehicle value is used for virtual 0 hour values

## PCP (pentachlorophenol) dataset

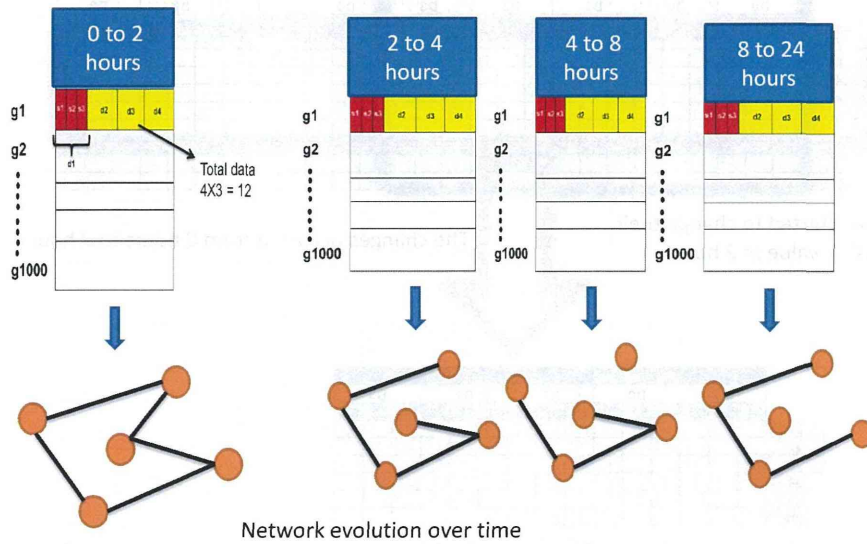
- 98, 55, 127, and 1192 genes started to change in response to PCP treatment at 2, 4, 8, and 24 hours.
- By using the list of genes and time-course of the genes, we will generate a gene network at each time points.

## Schematic Example

7

For a pertubagen (drug/chemical): P

For a tissue: T



## Inference of a network at 0 – 2 hours

8

	T0			T1									T2																				
	D0			D0			D1			D2			D3			D0			D1			D2			D3								
	S1	S2	D3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3			
g1																																	
g2																																	
g3																																	
g4																																	
g5																																	
.																																	
.																																	
g20000																																	

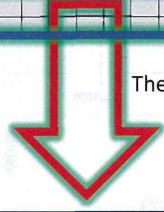


# Inference of a network at 0 – 2 hours

	T0			T1						T2												
	D0			D0			D1			D2			D3	D0			D1	D2		D3		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	
g1																						
g2																						
g3																						
g4																						
g5																						
...																						
g20000																						

94 genes started to change their expression value at 2 hours

The changes occurred from 0 hours to 2 hours.



	T0			T1												
	D0			D0			D1			D2			D3			
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	
g1																
g2																
g3																
g4																
g5																
...																
g94																

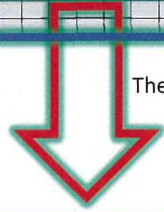
Expression values of 94 genes at 0 – 2 hours

# Inference of a network at 0 – 2 hours

	T0			T1						T2												
	D0			D0			D1			D2			D3	D0			D1	D2		D3		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	
g1																						
g2																						
g3																						
g4																						
g5																						
...																						
g20000																						

94 genes started to change their expression value at 2 hours

The changes occurred from 0 hours to 2 hours.



	T0			T1												
	D0			D0			D1			D2			D3			
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	
g1																
g2																
g3																
g4																
g5																
...																
g94																

Expression values of 94 genes at 0 – 2 hours

