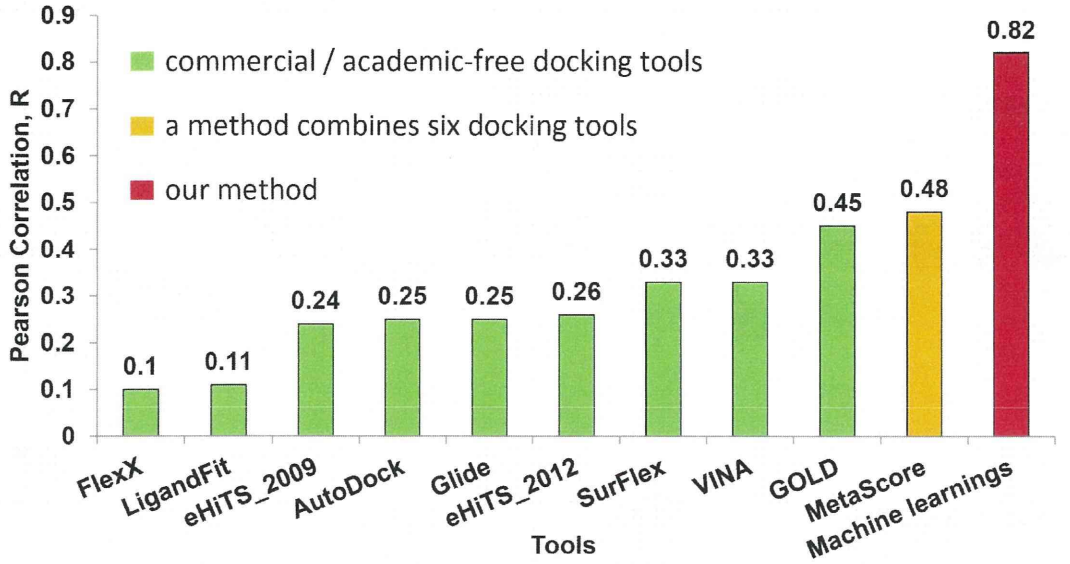
Performance of Our Method in Docking Simulation Compared with Others



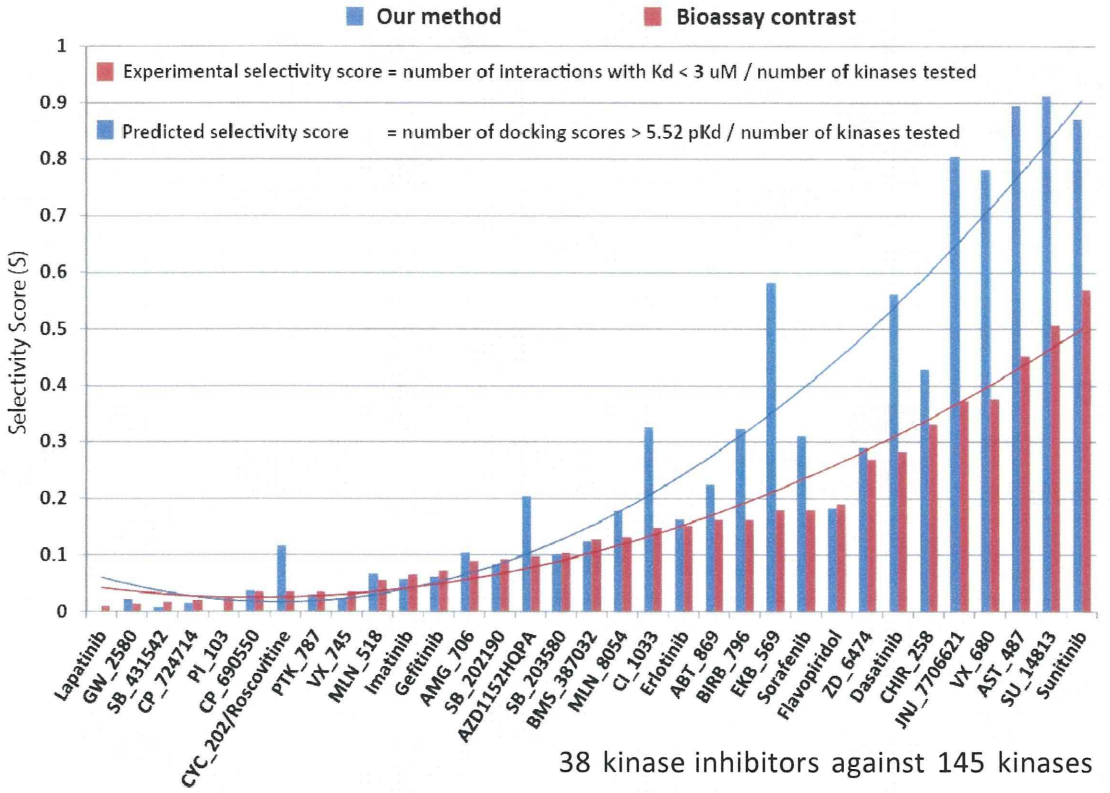
Our method:

- Multiple docking tools: eHiTS, GOLD and AutoDock VINA
- Machine learning systems
  - Machine learning system A: re-scoring function.
  - Machine learning system B: binding mode selection function.
- Hsin, K.Y., Ghosh, S. & Kitano, H. Combining Machine Learning Systems and Multiple Docking Simulation Packages to Improve Docking Prediction Reliability for Network Pharmacology. PLoS One 8, e83922 (2013).

5

Prediction of Inhibitor Selectivity Using Our Method Compared with Bioassay Contrast

(Consistent tendency with bioassay contrast)



38 kinase inhibitors against 145 kinases

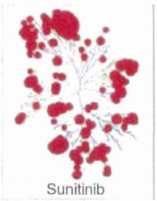
6

Case Study Using Our Method over MAPK Pathway

(Good consistency with bioassay contrast)

Bioassay contrast

Bioassay contrast

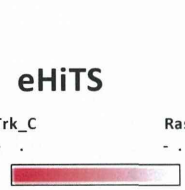


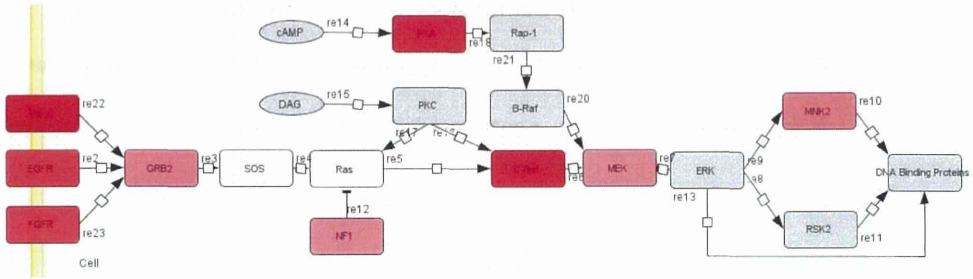
Case Study of Lapatinib Applying Tools Commonly Used in Pharmaceutical Industry

(Poor consistency with bioassay contrast)

eHiTS

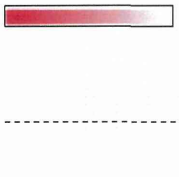
Trk_C Ras

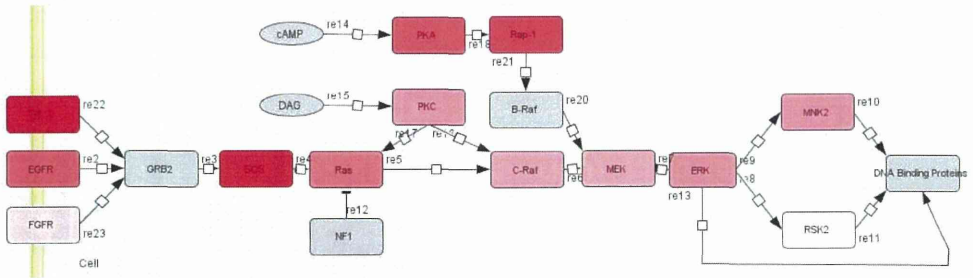




GOLD

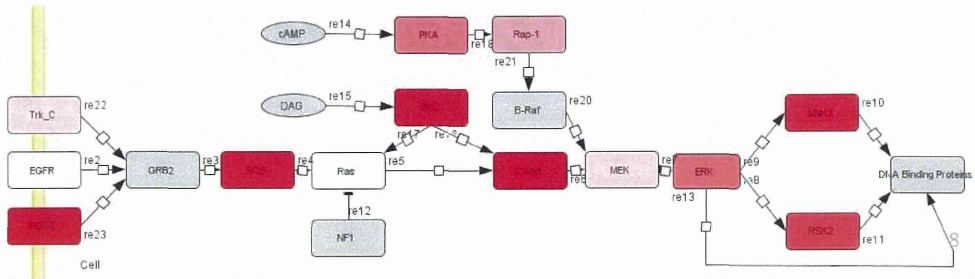
Trk_C RSK2





AutoDock VINA

cRaf Ras



Creating Next Generation Computational Tox Screening System

9

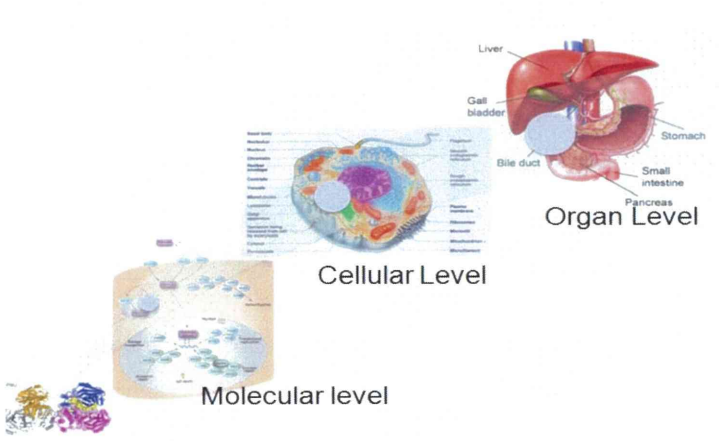
Toxicogenomics Annual Meeting

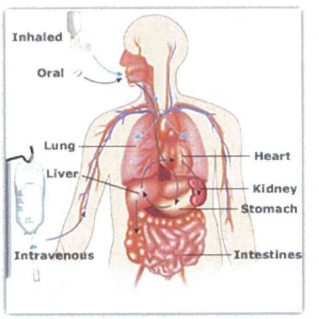
Gene expression and transcription regulation analysis for gene network discovery

13/02/2015

Systems Biology Institutee
Natalia Polouliakh, Hiroaki Kitano

Data-driven analytic pipeline



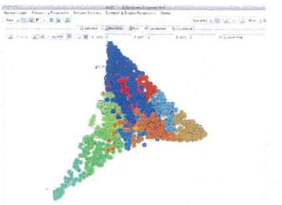


DNA chip

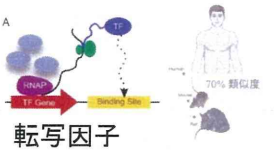
Gene Expression and Regulation analysis → PPI Pathway Map Simulation Model → Physiological Model Simulation Model

Research Agenda

- **AGCT** A Geometric Clustering Tool
  - Clustering Percellome data based on similarity of gene expression profile. Application to TCDD and TCDF (2,3,7,8-Tetrachlorodibenzo-p-dioxin and 2,3,7,8-Tetrafuran) chemicals.
    - Sample normalization
    - Dimension reduction on spectral manifold
    - Unsupervised clustering



- **SHOE** Sequence Homology in Higher Eukaryote
  - Phylogenetic footprinting for discovery of transcription regulation network



転写因子

3

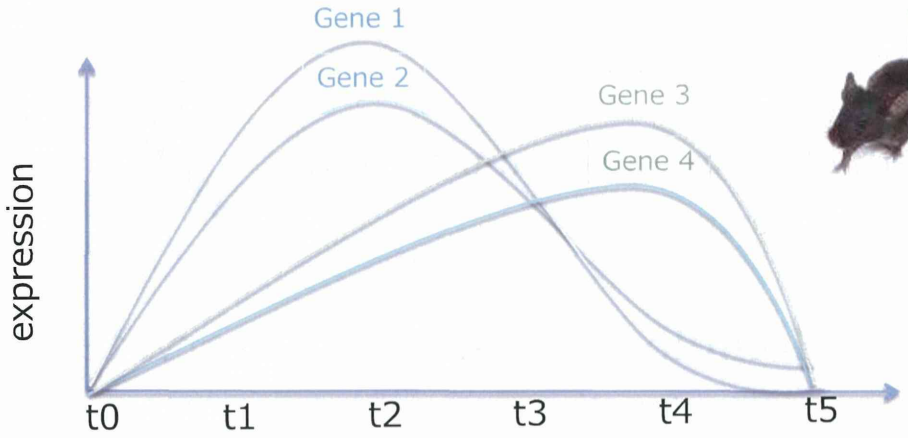
AGCT evolution

- AGCT 1.0 Interface 1.0 (2008) Richard Nock (Univ. of Martinique), Frank Nielsen (Sony CSL)
  - Small data (2,000-3,000 genes)
    - (published in 2009 Polouliakh et al, PLoS One)
- AGCT 2.0 Interface 1.0 (2010-2011)
  - Algorithm significantly improved (40,000*n)≈2hours
  - Source code partly re-factored
    - (preparing for submission)
- AGCT 3.0 Interface 2.0 (2012-2013)
  - Interface format changed
  - New Visualization added
  - Results Validation added
  - Source code rewritten in a module-like manner
  - Plugin to CellDesigner
  - Plug-in to Garuda platform is developed

4

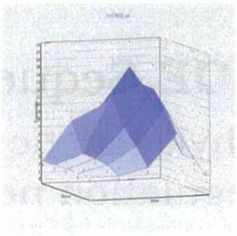
Atlas of Cell Life by AGCT

- AGCT reconstructs gene network basing on the similarities of the expression profiles of genes



- chemical
- doze
- time
- expression

40,000 probes

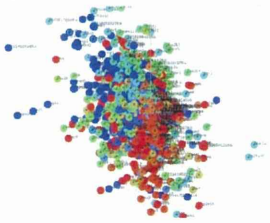


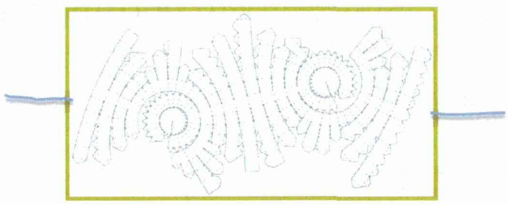
5

Processing data on AGCT

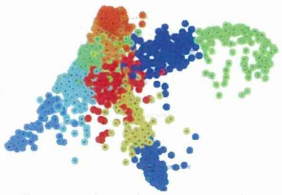
1. 時系列データ前処理:線形回帰/ウェーブレット変換
2. 遺伝子間の類似度マットリックス
3. 低次元に落とすためにSpectral clusteringを行う。通常の主成分分析も行う。
4. 発見的なClustering法を使って構造上でデータの分割を行う。
5. 結果のinteractive visualizationやscenario 記録を行う。

PCA : M*N matrix





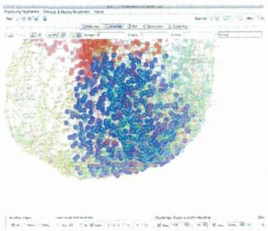
Spectral clustering:
M*M matrix



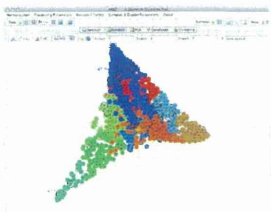
Orthogonal Laplacian matrix to compute one dimension per cluster/gene

Examples of different network topologies

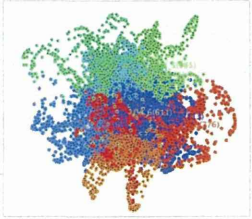
Mouse Stem cell



TCDD affected mouse liver cell

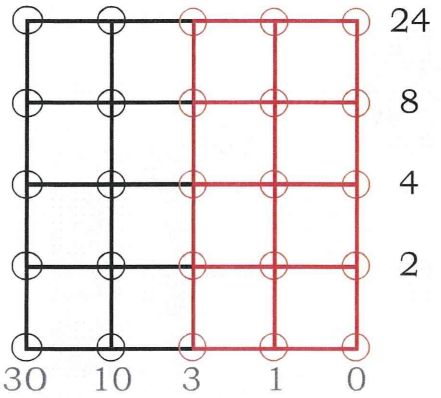


Influenza affected mouse bronchi cell

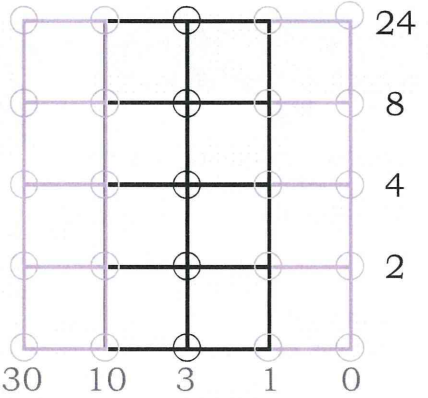


6

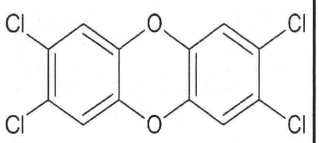
TCDD



TCDF



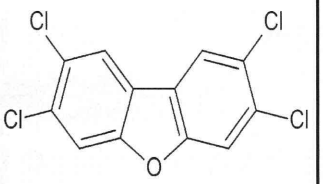
TTG020
TCDD
(2,3,7,8-Tetrachlorodibenzo-p-Dioxin)
CAS 1746-01-6
MW=321.9744



vehicle[corn oil]
dosage[0,1,3,10,30microg/kg]
time[2,4,8,24hr]

TEF : 0.1

TTG026
TCDF
(2, 3, 7, 8-Tetrachlorodibenzofuran)
CAS 51207-31-9
MW=305.975

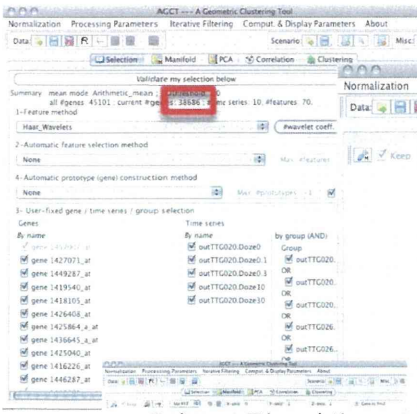


vehicle[corn oil]
dosage[0,1,3,10,30microg/kg]
time[2,4,8,24hr]

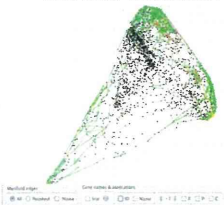
7

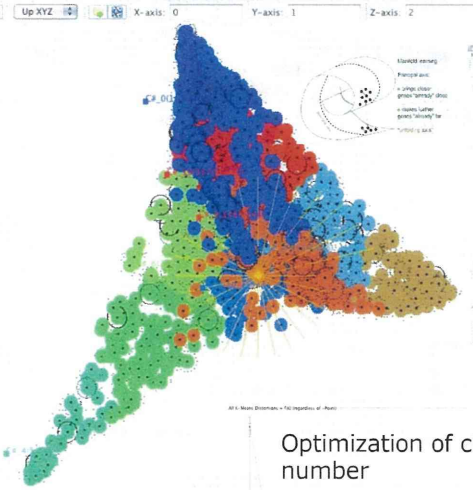
Analysis of TCDD-TCDF **together** data (≈40,000) Plug-in to CellDesigner to produce gene MAP

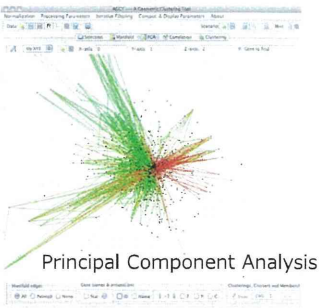
Clustering: Bergman k-means
Affinity propagation, Expectation Maximization, CP, NNMF.



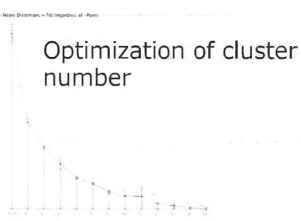
Delaunay Triangulation
Surface Extraction (p≤0.05)







Principal Component Analysis



Optimization of cluster number

Cluster centroid

8

Matrix: 5 doze * 4 time * 3 replicate * 40,000=2,400,000^3 point = NP hard

Optimization

Matrix sparsfication

m eigenVectors O(n-genes d=10 m=100) where d density for sparsification

(1) Sparse similarity matrix

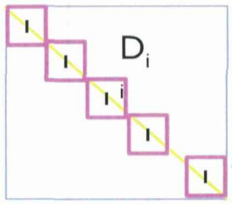
W =

$d_{ij} \ldots d_{ij2} \ldots d_{ij10}$

$O(n^2+10)$ instead of $O(n^3)$

(2) Laplacian matrix

L =



$L = D^{-1/2}WD^{-1/2}$

$D_{ii}^{-1/2} = 1/\mathrm{sqrt}D_{ii}$

(3) Lanczos algorithm

(Transforms the original matrix into a tridiagonal matrix which is real and symmetric.)

Note that (x,y) represents the dot product of vectors x and y here.

After the iteration, we get the α_j and β_j which constructs a tridiagonal matrix

$$T_{mm} = \begin{pmatrix} \alpha_1 & \beta_2 & & & 0 \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_{m-1} \quad \beta_m \\ 0 & & & & \beta_m \quad \alpha_m \end{pmatrix}$$

The vectors v_j (**Lanczos vectors**) generated on the fly constructs the transformatic

V_m

which

or coul

eps the first vector

Matrix W … constant factor speedup by memorization

I=

Q

QT = $QQ^TLQ=QT=LQ$

Q – orthogonal matrix

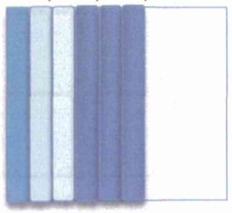
Q^T – transformed orthogonal matrix

L – normalized Laplacian

(4)

e1, e2, e3,..

M =



~100

12 hours!

160万倍計算UP!

O(n3) -> O(d*m*n)

d=10 m=100

n-probes

$Lq_1 = \alpha_1 q_1 + \beta_1 q_2;$

$Lq_2 = \beta_2 q_1 + \alpha_2 q_2 + \beta_3 q_3;$

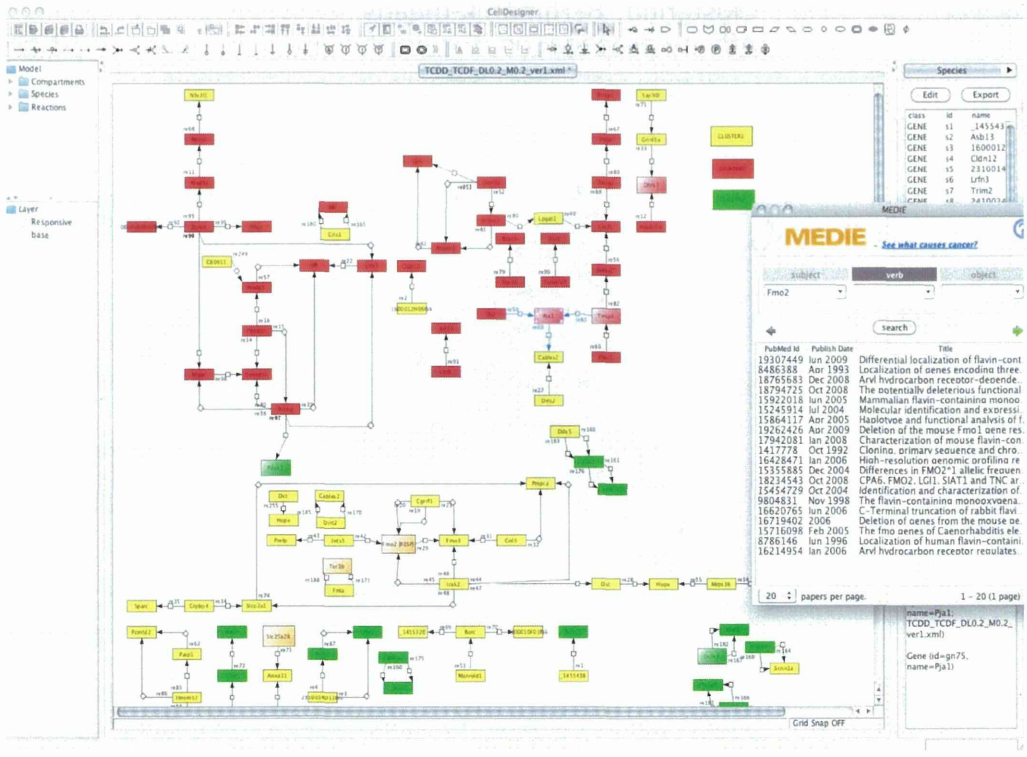
$Lq_m = \beta_m q_m + \alpha_m \beta_m;$

The degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

Degree matrix D is defined as the diagonal matrix with the degrees $d_1, \ldots d_n$ on the diagonal.

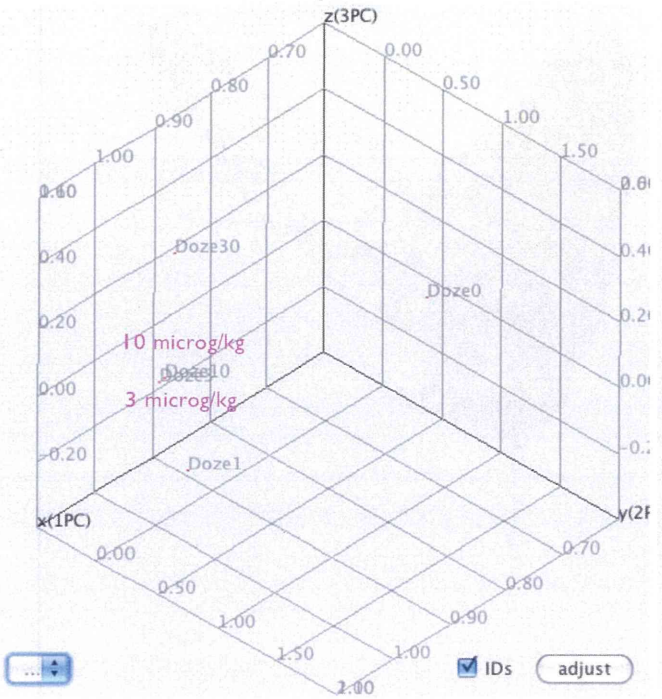
Plugin to CellDesigner -> MAP



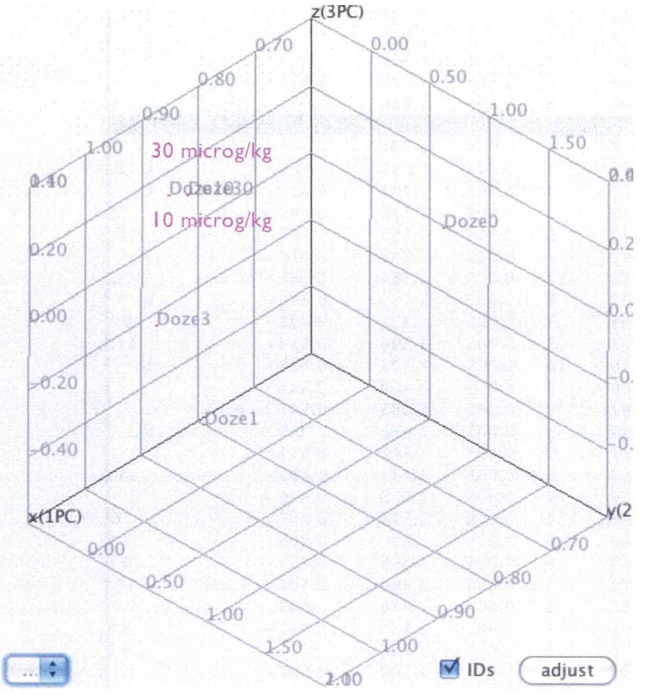
10

AGCT3.0: TCDD and TCDF: PCA on 5 doses

7,881 probes

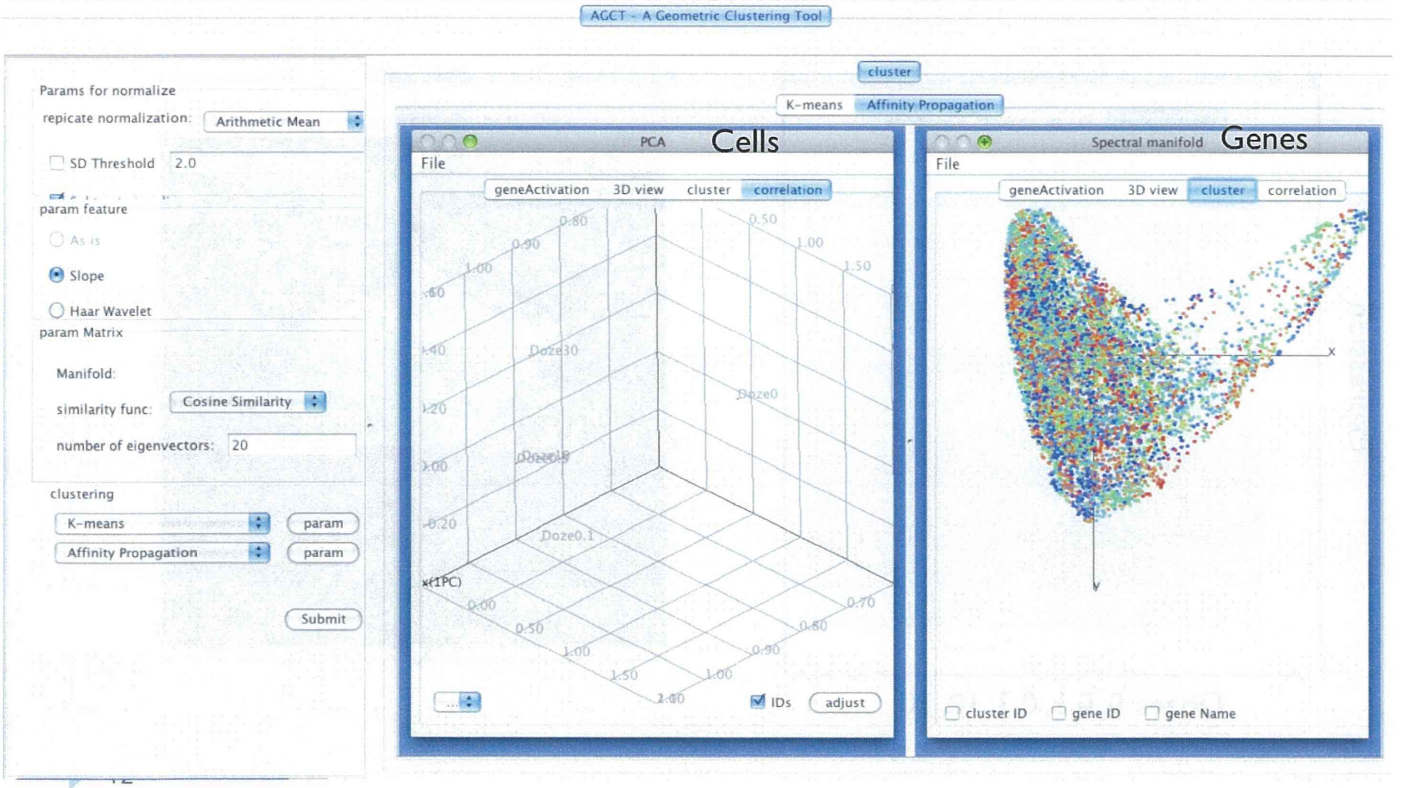


8,674 probes



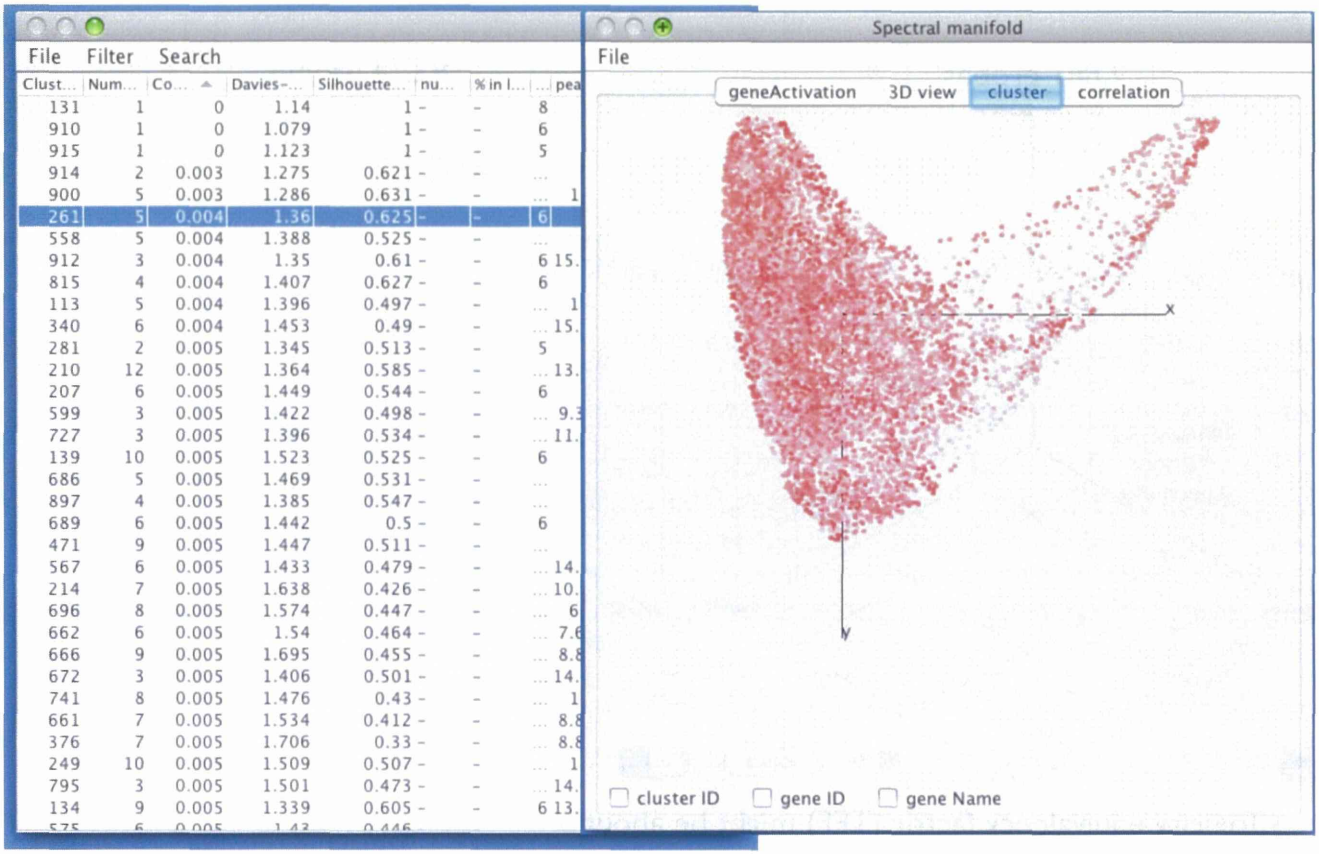
- Toxicity equivalency factor (TEF) might be about 3.

TCDD spectral clustering view

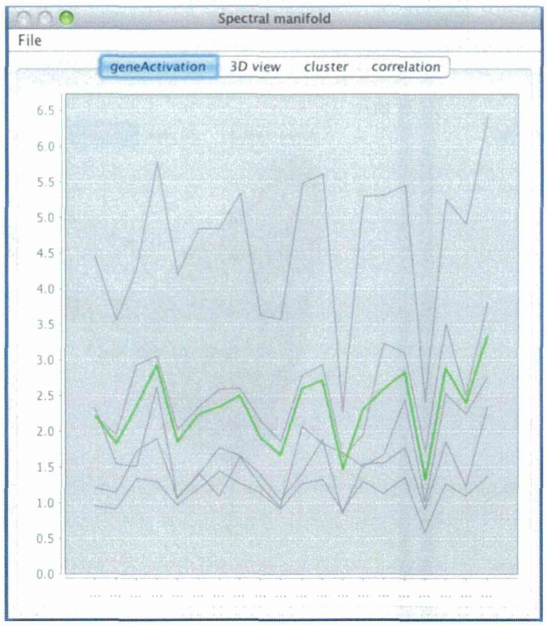


CLUSTER SORTING BY VALIDITY

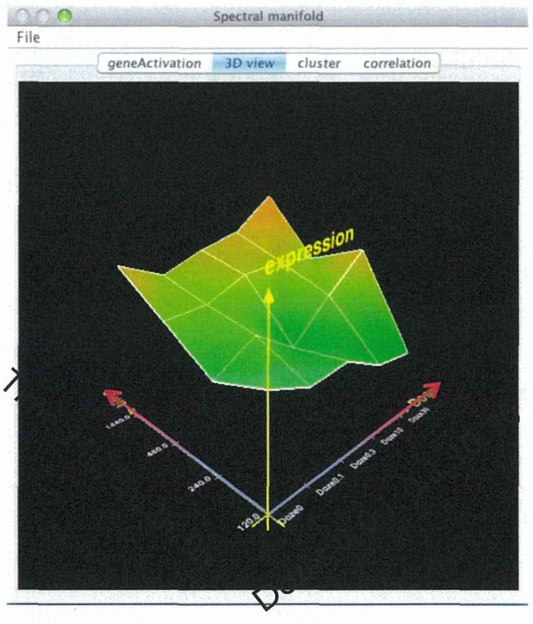
Compactness, Davies-Bolduin index, Sihouette index, Peaks,



2D and 3D visualization of clusters

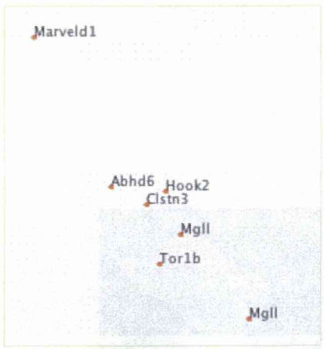


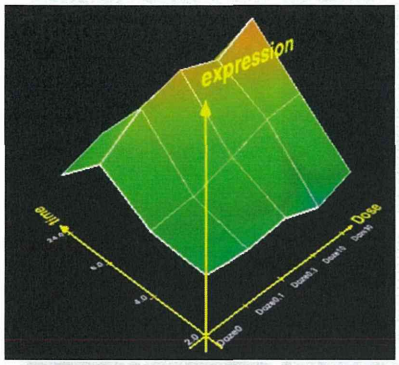
Dozes 0, 0.1, 0.3, 10, 30

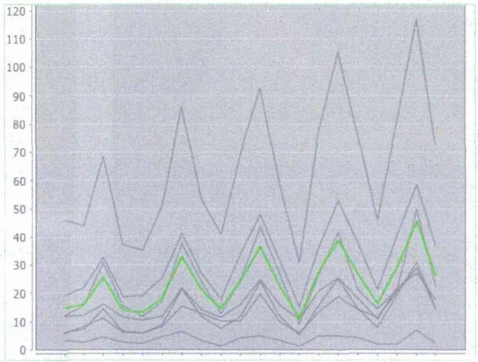


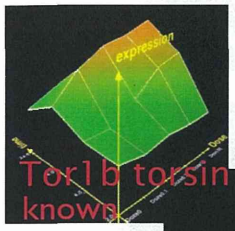
Cluster examples on TCDD

7,881 genes (1661cl) -> 3,117 probes in (498 cl) 4,550 (60%) *unrelated*



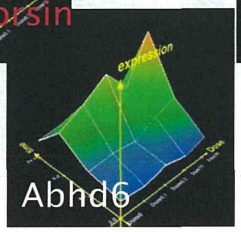




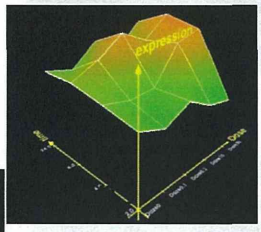


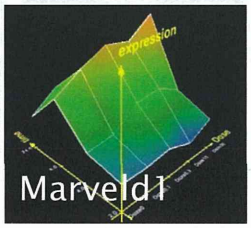
Tor1b torsin known

Immune-Mediated disorder

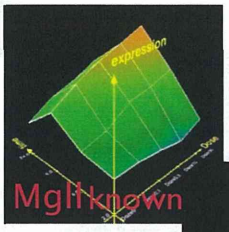


Abhd6



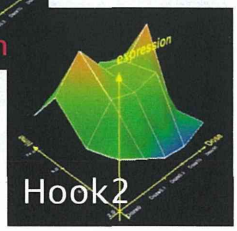


Marveld1



Mgll known

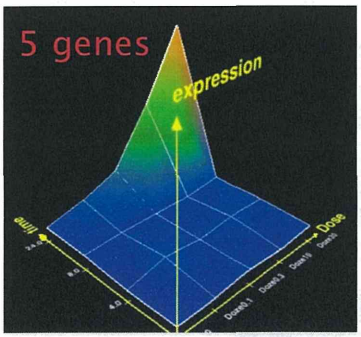
Hydrolyze Fatty acids

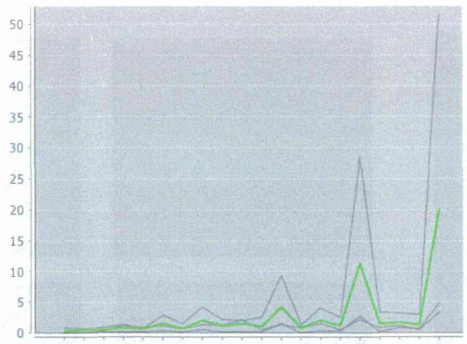


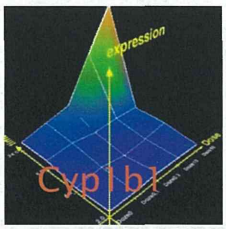
Hook2

15

TCDD Ahr cluster

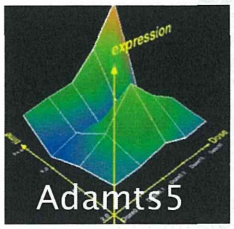






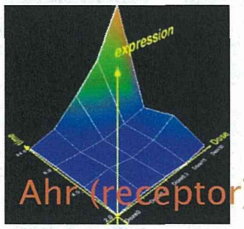
Cyp1b1

Metabolic enzyme

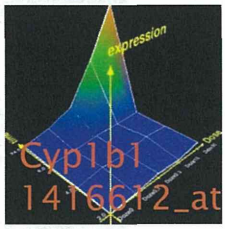


Adamts5

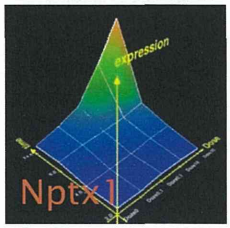
metalloproteinase



Ahr (receptor)



Cyp1b1 1416612_at



Nptx1

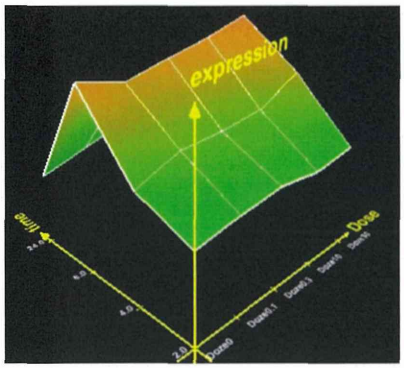
binding protein for the snake venom toxin taipoxin

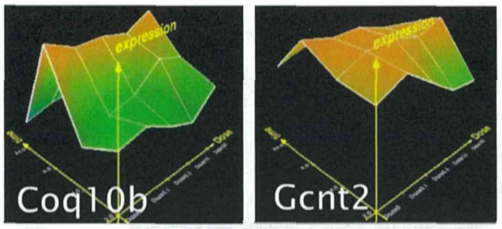
16

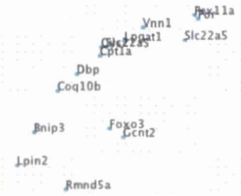
TCDD Dbp cluster

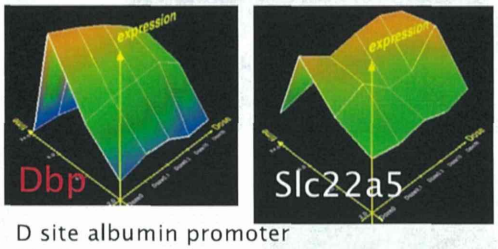
TF binding to Insulin gene (disease associated aging)

D site of albumin promoter (albumin D-box) binding protein

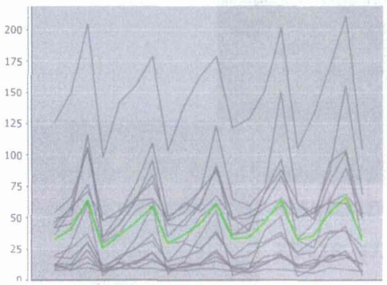


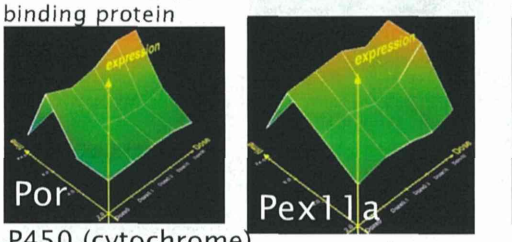






D site albumin promoter binding protein





P450 (cytochrome) oxidoreductase

peroxisomal biogenesis factor

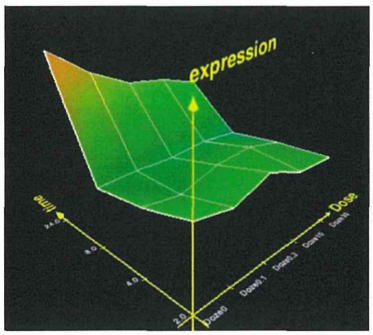
Lysophosphatidylglycerol acyltransferase 1

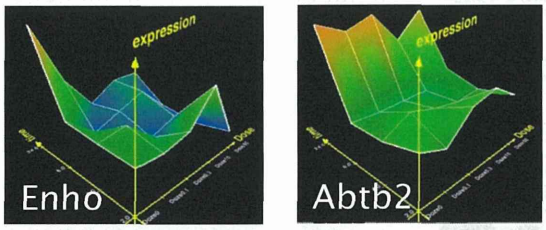
17

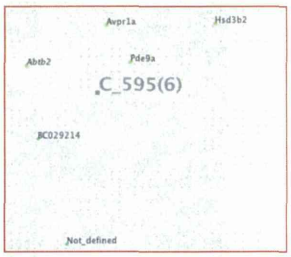
TCDD Hsd3b2 cluster

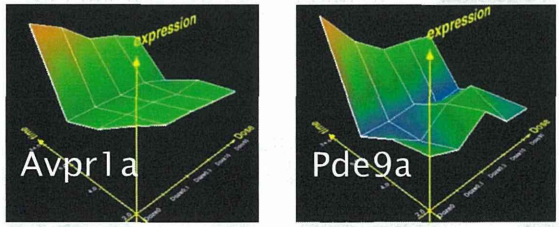
(steroid hormone production)

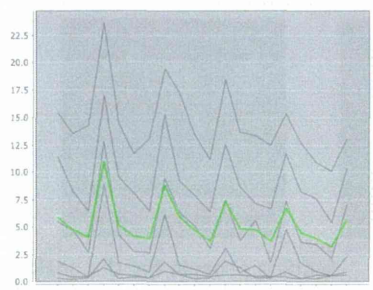
3beta-hydroxysteroid dehydrogenase/delta(5)-delta(4)isomerase type II

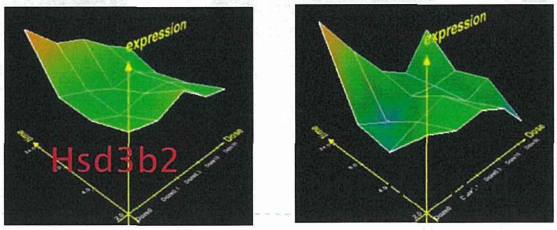










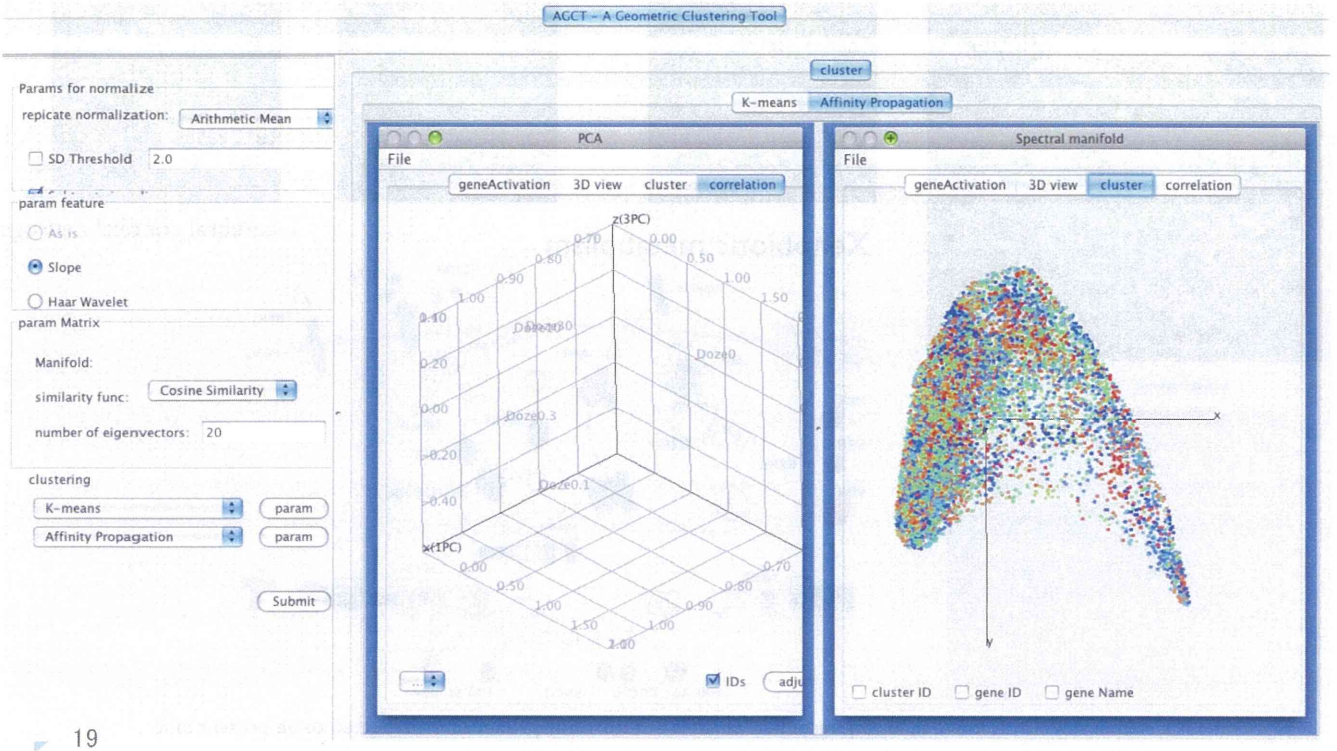


18

Hydroxy delta 5 steroid dehydrogenase 2

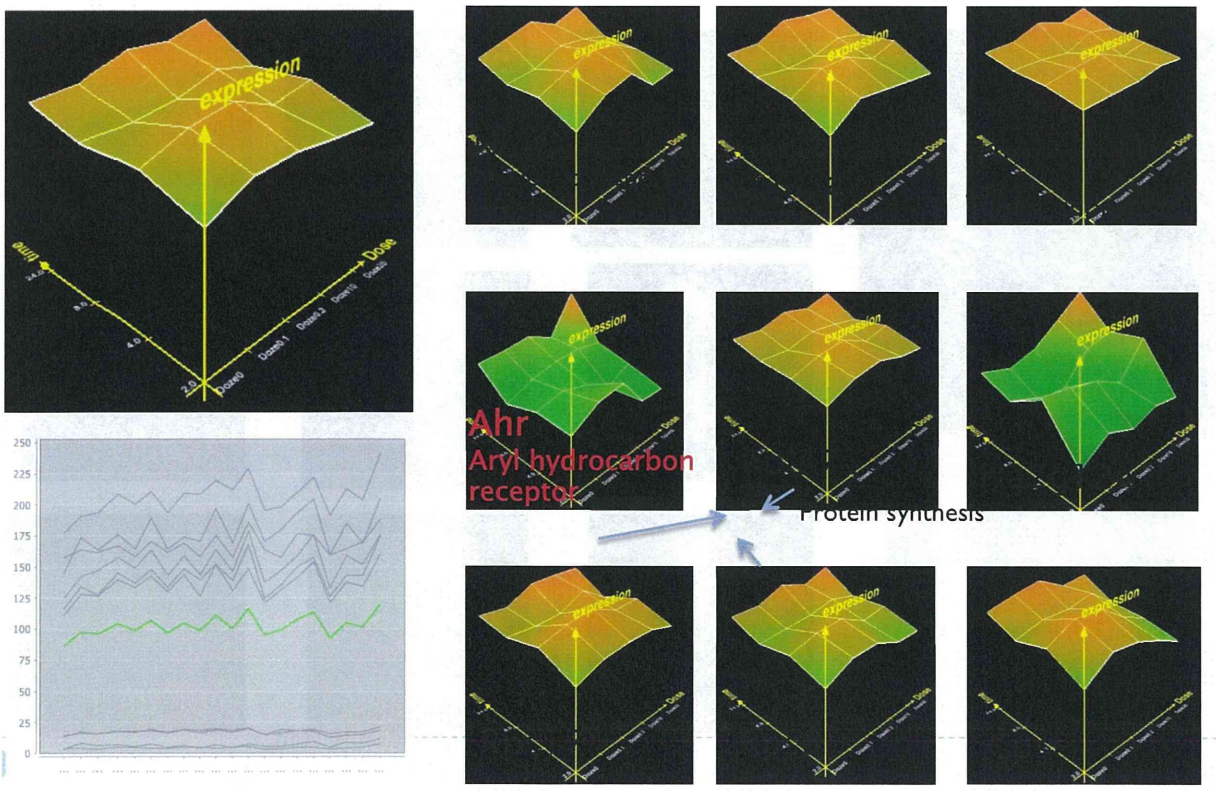
Cluster examples on TCDF

8,674 genes (1600cl) -> 3,771 probes in (369 cl) 4,903 (56%) *unrelated*

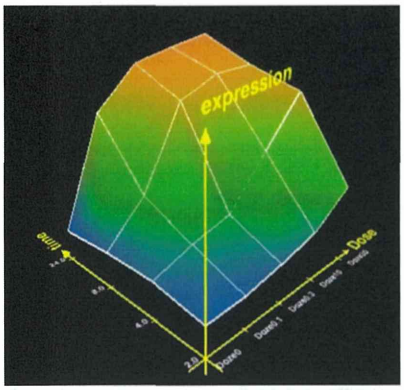


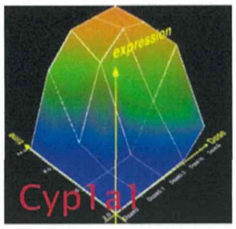
19

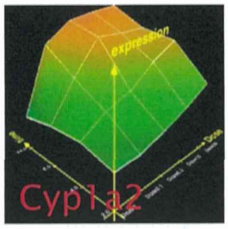
TCDF Ahr cluster

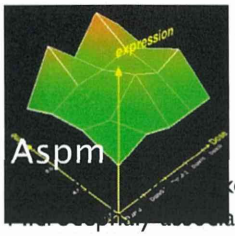


TCDF Cyp1a1 cytochrome P450

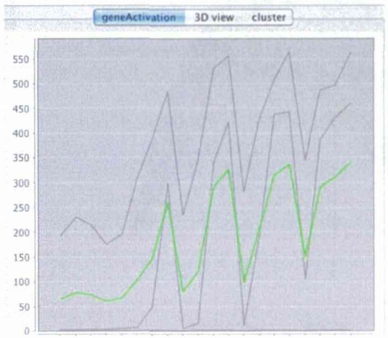




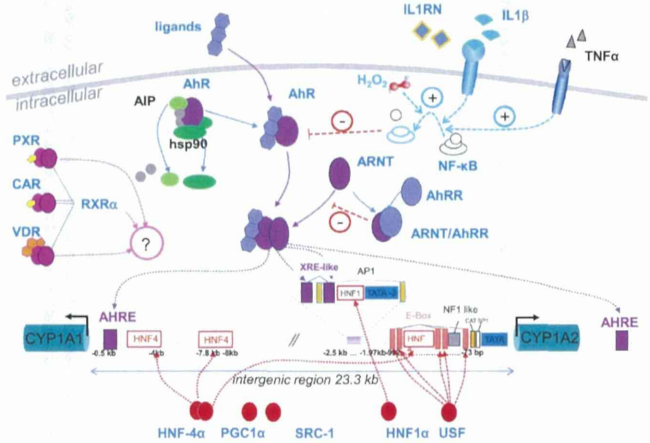




ke,
ated
cerebral cortical neurogenesis

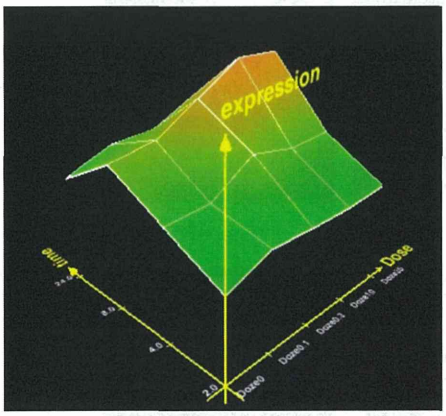


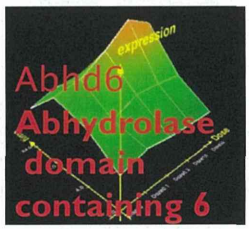
Xenobiotic metabolism

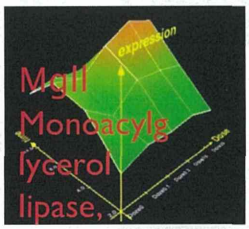


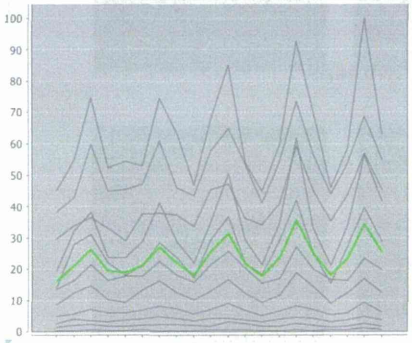
A xenobiotic is a chemical which is found in an organism but which is not normally produced or expected to be present in it. It can also cover substances which are present in much higher concentrations than are usual.

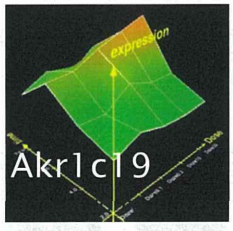
TCDF Abhd6 cluster



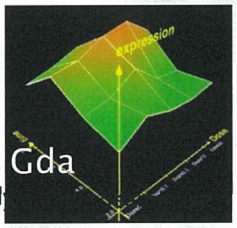


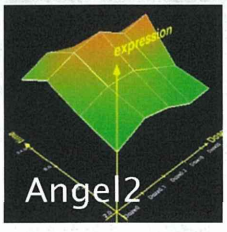


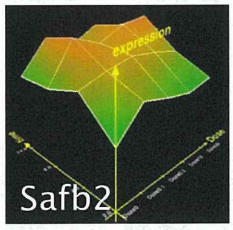


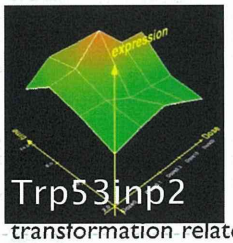


famil

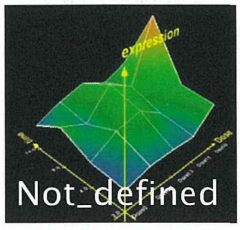




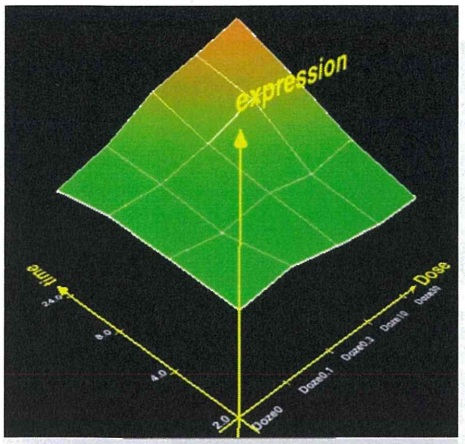


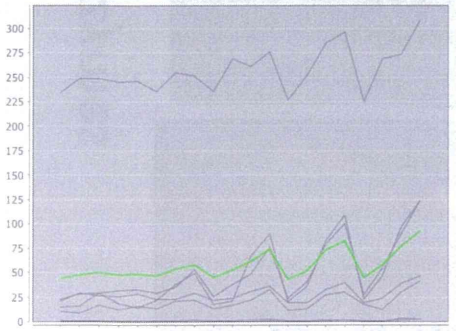


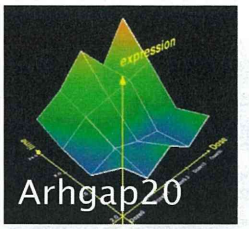
transformation related
protein 53

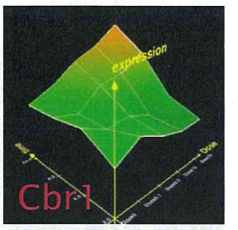


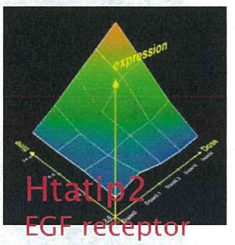
TCDF Slc46a3 cluster

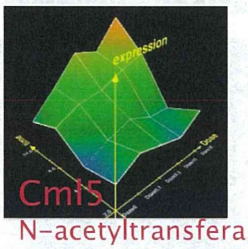


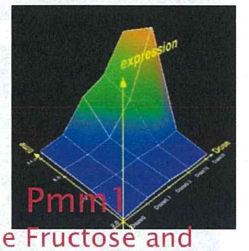




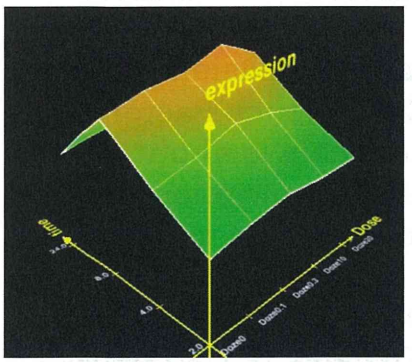


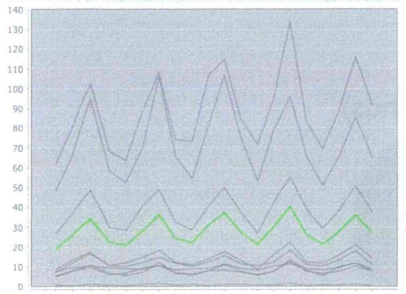


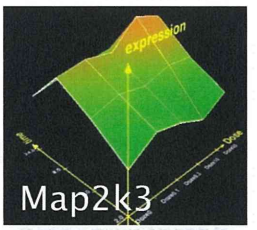


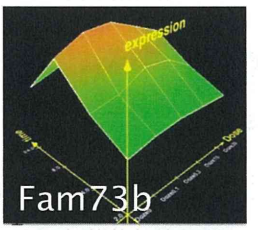


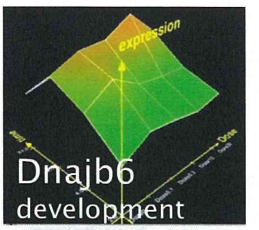
TCDF Ptplad1 cluster

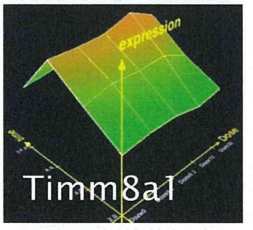


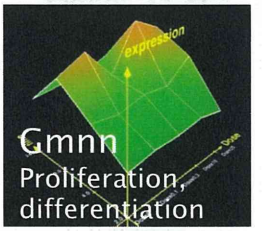


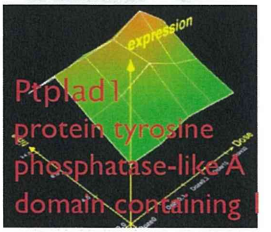


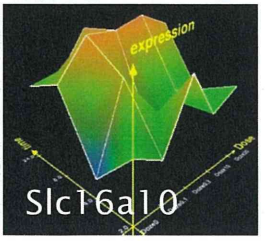


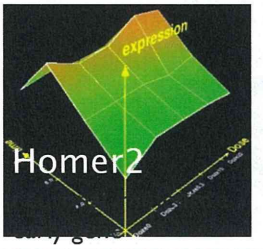


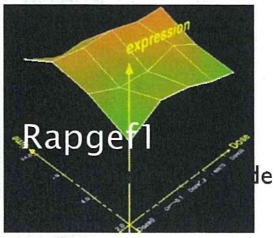






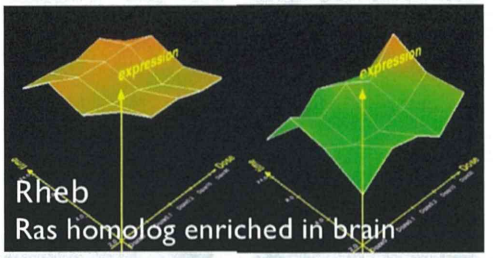


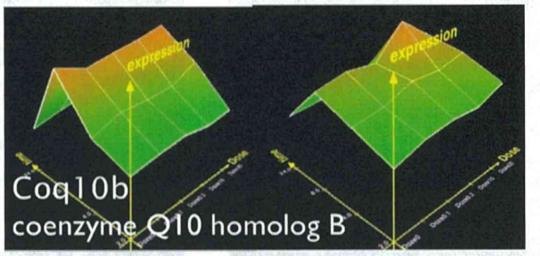


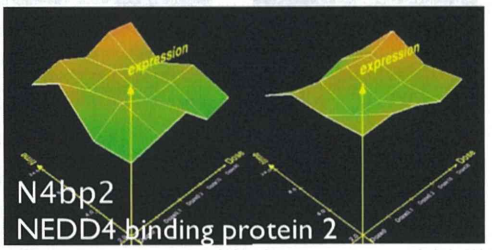


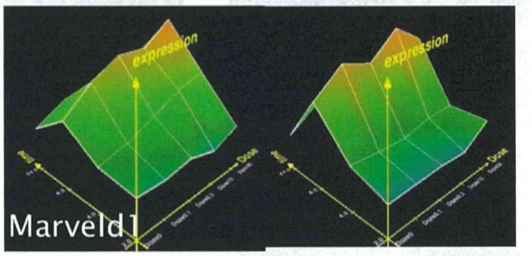
24

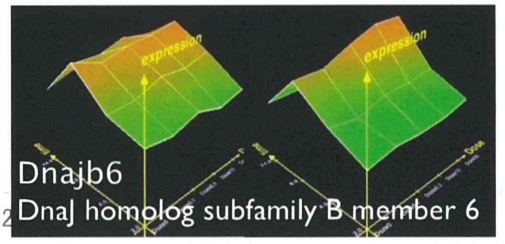
New genes shared regulated by TCDD and TCDF chemicals (600)

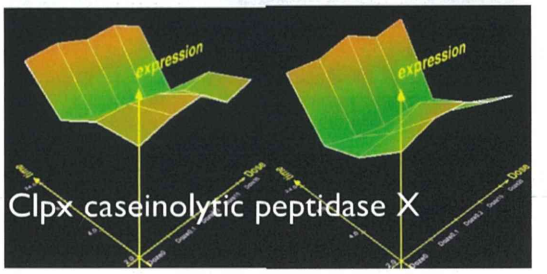










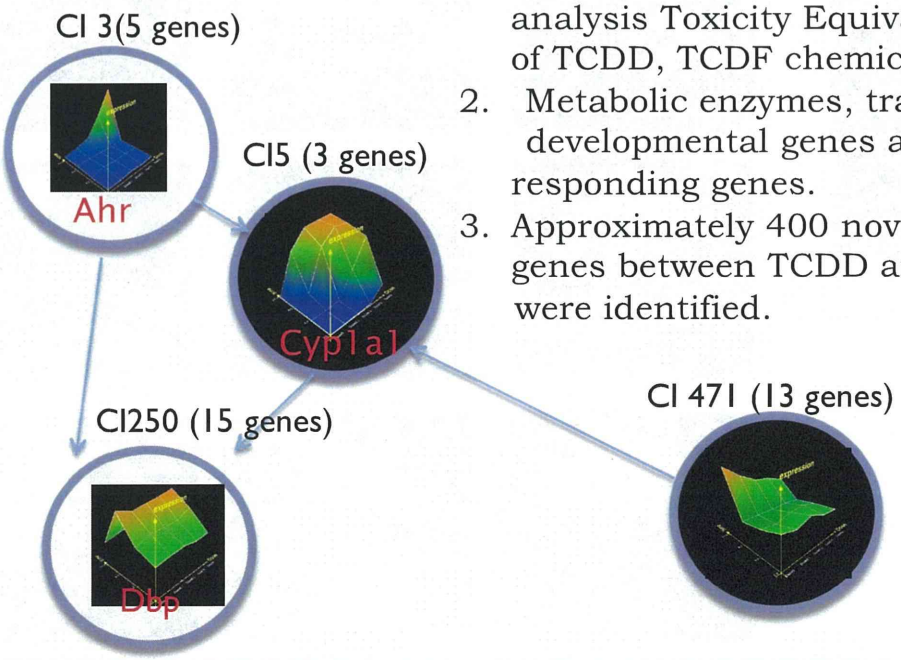


Metabolic enzymes

2

Conclusions

- Cluster network



1. Basing on whole data expression profile analysis Toxicity Equivalency Factor (TEF) of TCDD, TCDF chemicals is about 3.
2. Metabolic enzymes, transcription factors, developmental genes are the main responding genes.
3. Approximately 400 novel co-regulated genes between TCDD and TCDF chemicals were identified.

26

TCDD responsive probes (211) were used to map on CellDesigner

(Report on the analysis is attached)

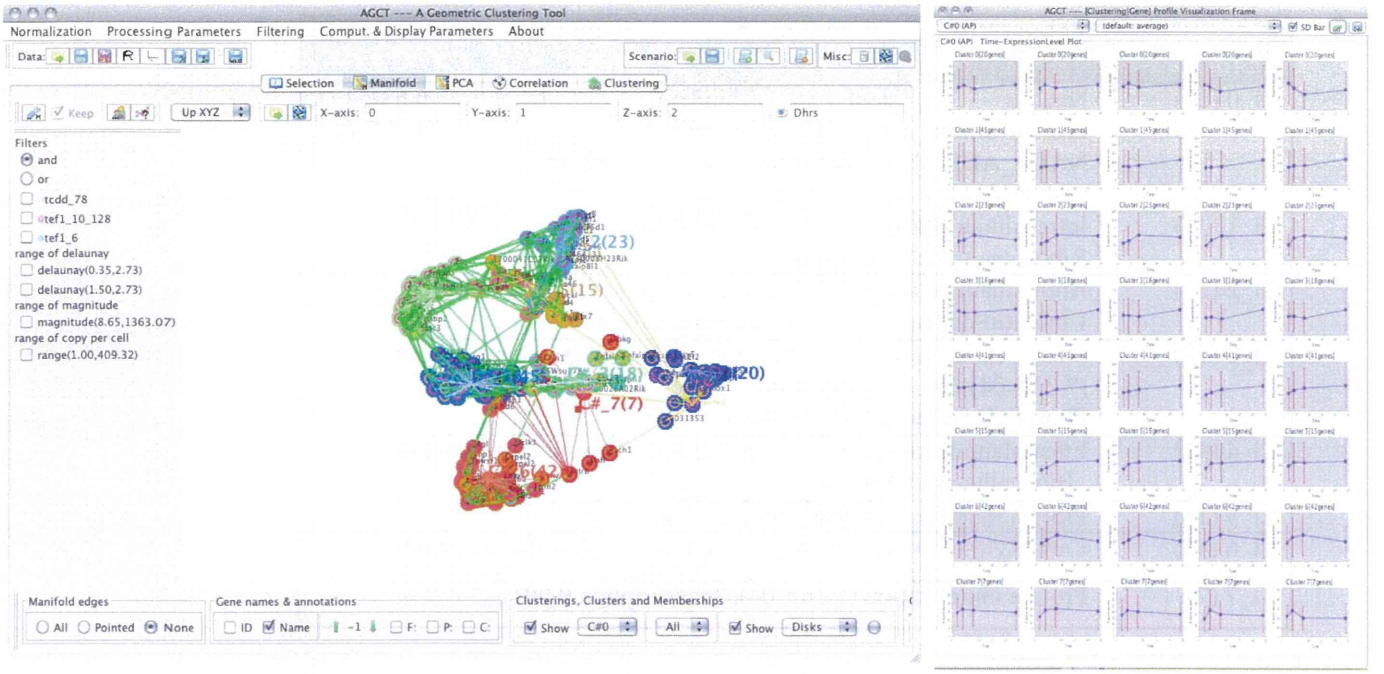


Figure 1: Eight clusters were were obtained by Affinity propagation method on 211 TCDD probes.

27

Gene ontology for 8 TCDD clusters

C2 antioxidant activity (0.001)
ubiquitin-protein ligase activity
ubiquitin-dependent protein catabolism

C4 Ethanolamine kinase activity 5.13E-04
Phosphatidylethanolamine biosynthesis

C1 intracellular (0.01)
Glutathione transferase activity
ribosome
translation

C5 protein modification_process 1.74E-04

C3 mast_cell_activation 4.11E-11
ethanolaminephosphotransferase
*phosphotransferase_*
Phospholipid biosynthetic process

C7 protein dimerization 1.16E-04
transcription 0.001

C0 nucleic acid binding 6.63E-17
protein_amino_acid_ADP-ribosylation 1.09E-05
NAD+_ADP-ribosyltransferase activity 1.09E-05
vasculogenesis 1.09E-05

C6 Carboxylesterase 7.10E-05
Phosphatidate phosphatase

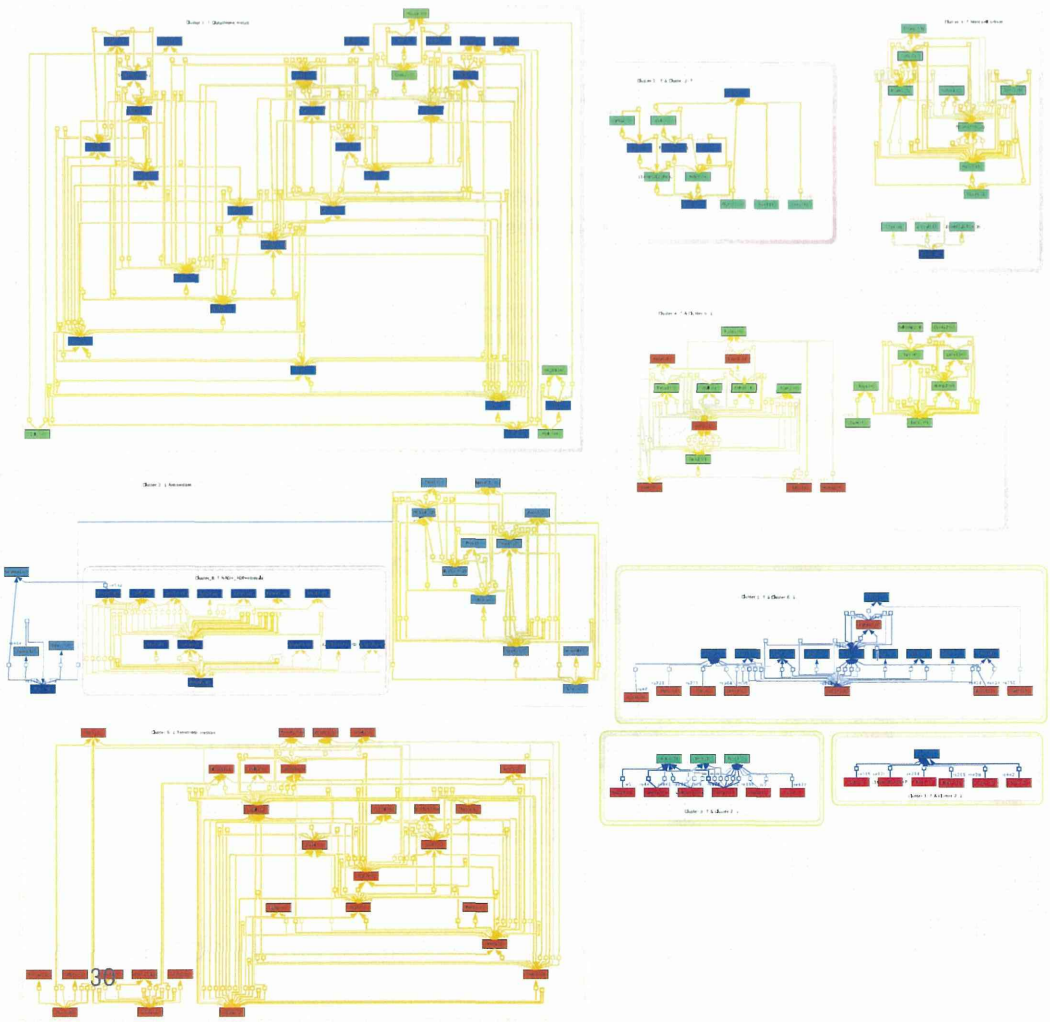
Figure 2: GO terms for eight clusters obtained by Affinity propagation method on 211 TCDD probes.

28

Table 1. Classification of top three overrepresented GO Biological Process categories in eight clusters of 211 probes for eight 2,3,7,8-tetrachlorodibenzodioxin TCDD clusters.

Cluster (probe)	GO term and p-value (chi2, Yates's correction)
Cl_0 (20)	Nucleic acid binding (6.63E-17), Protein_amino_acid_ADP-ribosylation (1.09E-05), NAD+_ADP-ribosyltransferase activity (1.09E-05), vasculogenesis (1.09E-05)
Cl_1 (45)	Intracellular (0.01), Glutathione transferase activity (0.01,) Ribosome translation (0.01)
Cl_2 (22)	Antioxidant activity (0.001), Ubiquitin-protein ligase activity (0.001), Ubiquitin-dependent protein catabolism (0.001)
Cl_3 (18)	Mast_cell_activation (4.11E-11), Ethanolaminephosphotransferase (7.22E-04), Phosphotransferase (7.22E-04), Phospholipid biosynthetic process (7.22E-04)
Cl_4 (41)	Ethanolamine kinase activity (5.13E-04), Phosphatidylethanolamine Biosynthesis (5.13E-04)
Cl_5 (15)	Protein modification_process (1.74E-04)
Cl_6 (42)	Carboxylesterase (7.10E-05), Phosphatidate phosphatase (0.001)
Cl_7 (7)	Protein dimerization (1.16E-04), Transcription 0.001

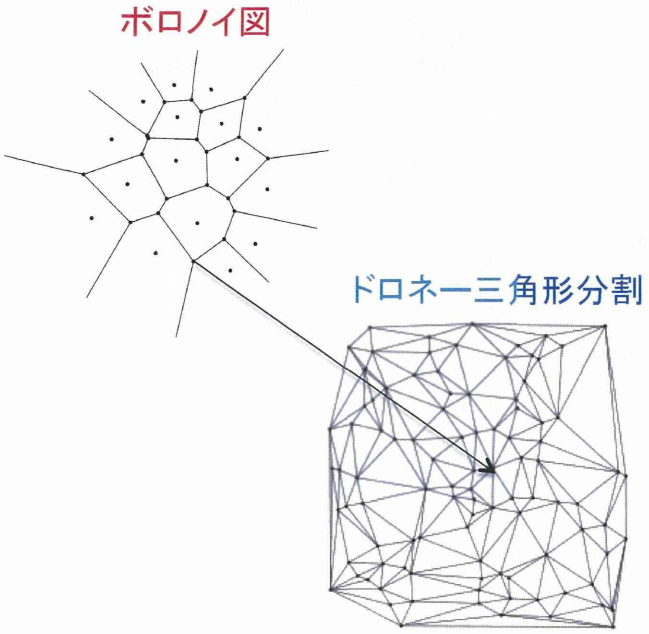
29



36

Delaunay Triangulationドロネー図

- 離散幾何学ロシアの数学者ボリス・ドロネーに由来する



- 与えられたボロノイ図から対応するドロネー図を作図するには、与えられたボロノイ図の各領域（ボロノイ領域）に一つずつの特定の点（母点）を選んで固定し、どの二つのボロノイ領域についても、それが隣接ボロノイ領域ならば母点同士を結び、隣接していない場合は二つの母点を結ばないという操作を行う。

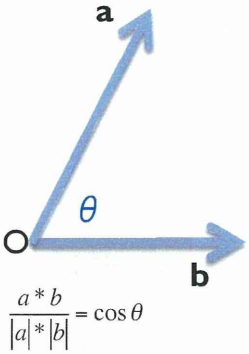
31

In AGCT DT is used for the validation of manifold

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:

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

Center of Manifold

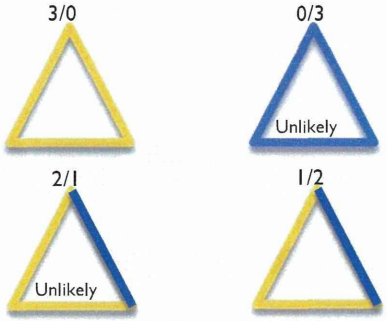


$$\frac{a*b}{|a|*|b|} = \cos\theta$$

if $\theta > \frac{\pi}{2}$

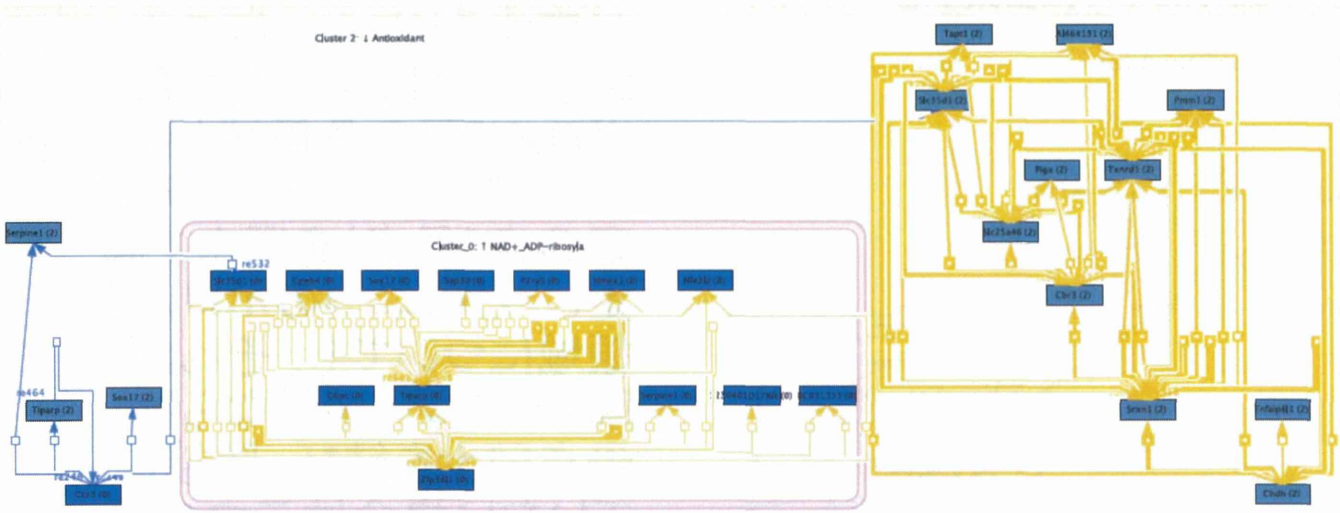
if $\theta < \frac{\pi}{2}$

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.



32

Drawn based on DT value



if $\theta > \frac{\pi}{2}$ Negative correlation

if $\theta < \frac{\pi}{2}$ Positive correlation