

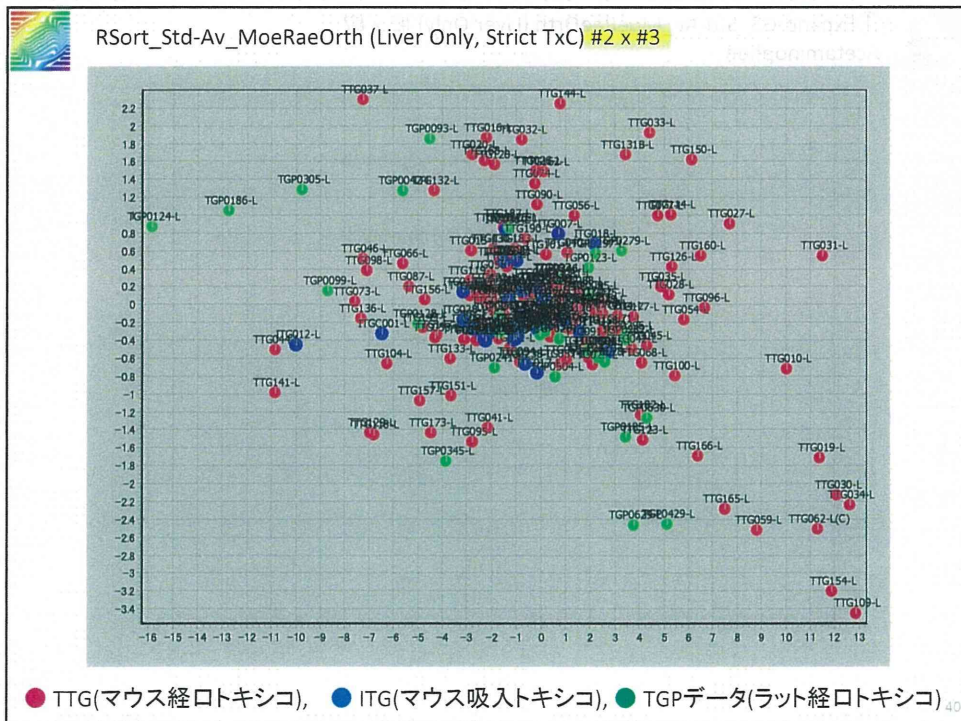
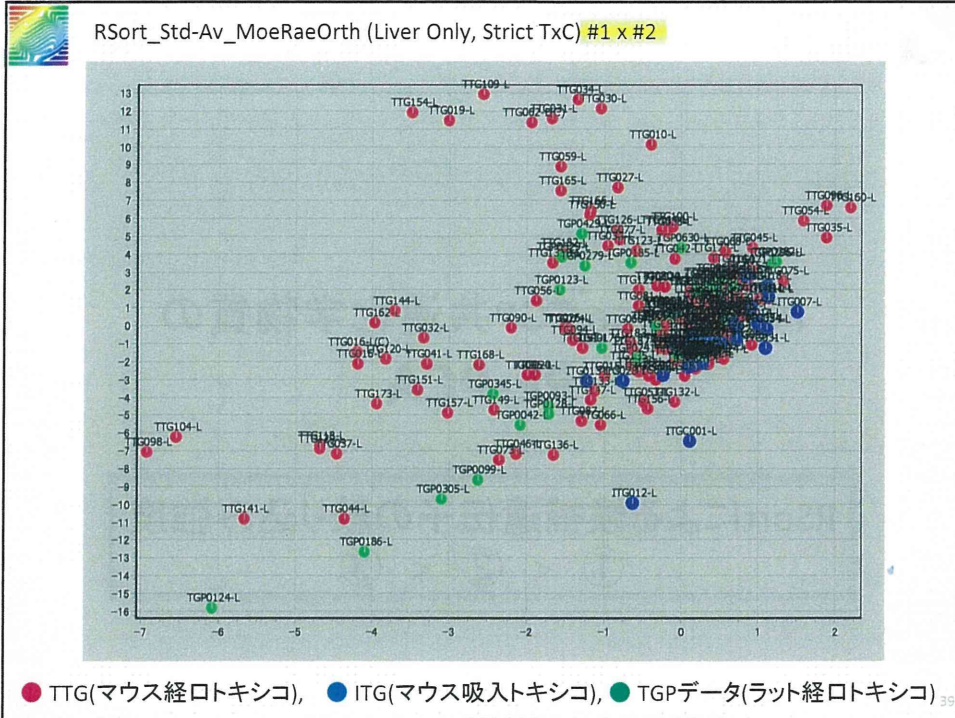
PE-PCA


③

RSort\_Std-Av\_MoeRaeOrth

Liver Only

Strict, Target c Candidate mode

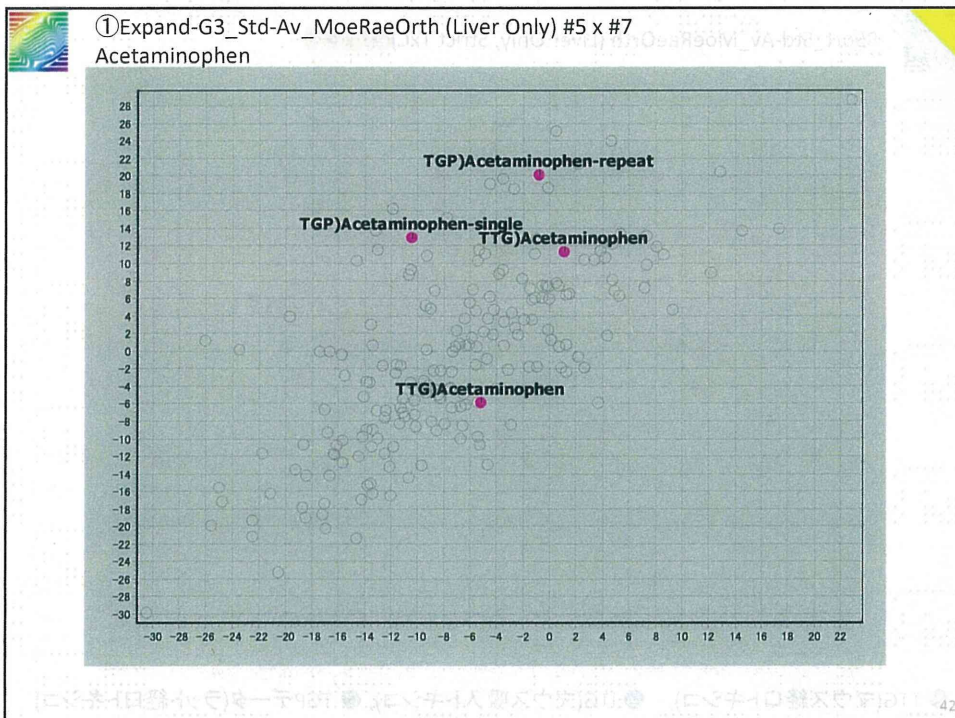




## Percellome-TGP共通化学物質の PE-PCA

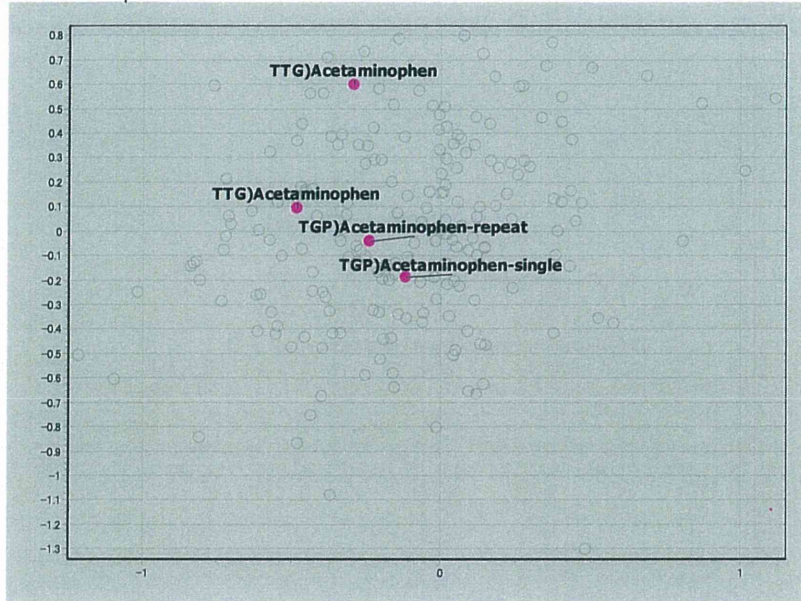
RSortによる候補遺伝子の絞り込み強度  
① < ② < ③

41





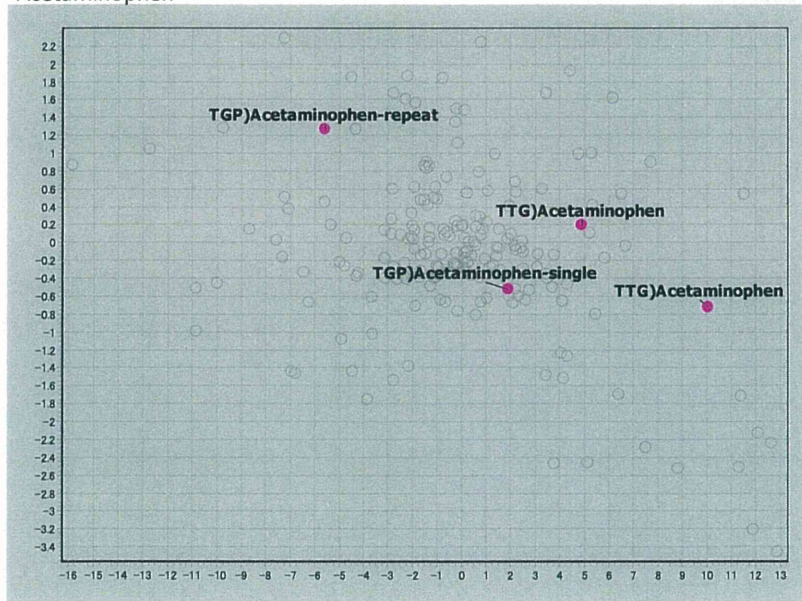
②Expand-G3\_Std-Av\_MoeRaeOrth (Liver Only, Strict TxC) #3 x #6  
Acetaminophen



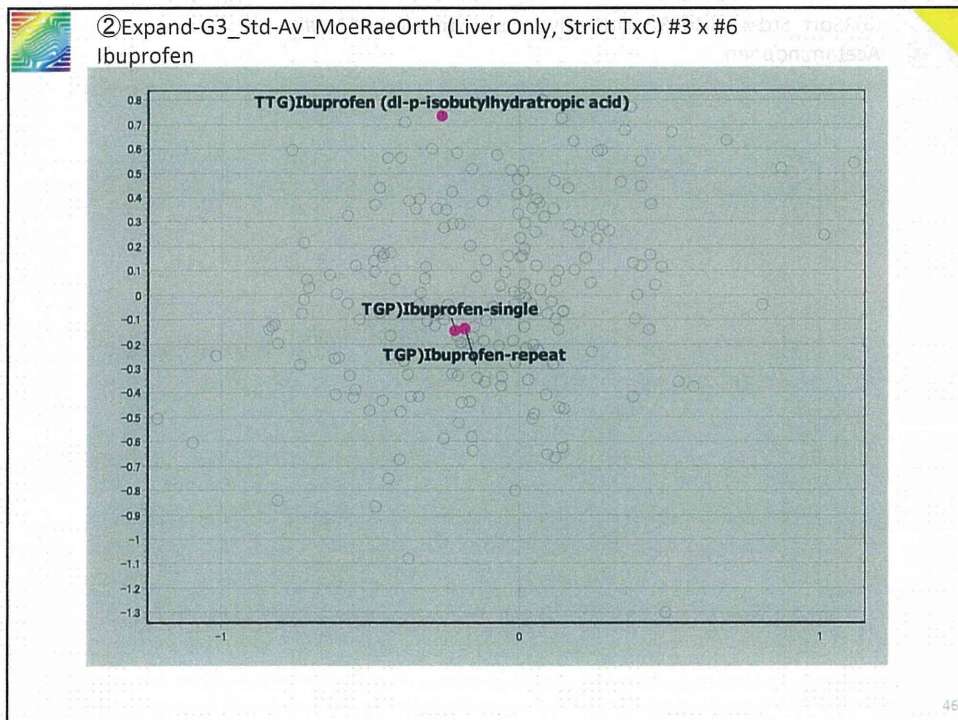
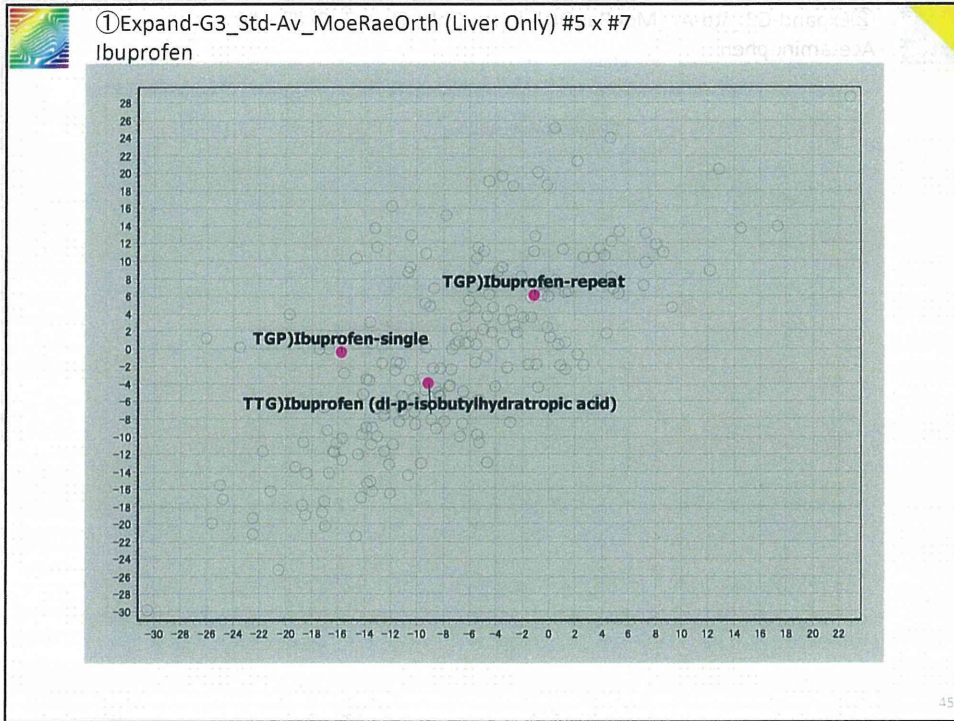
43

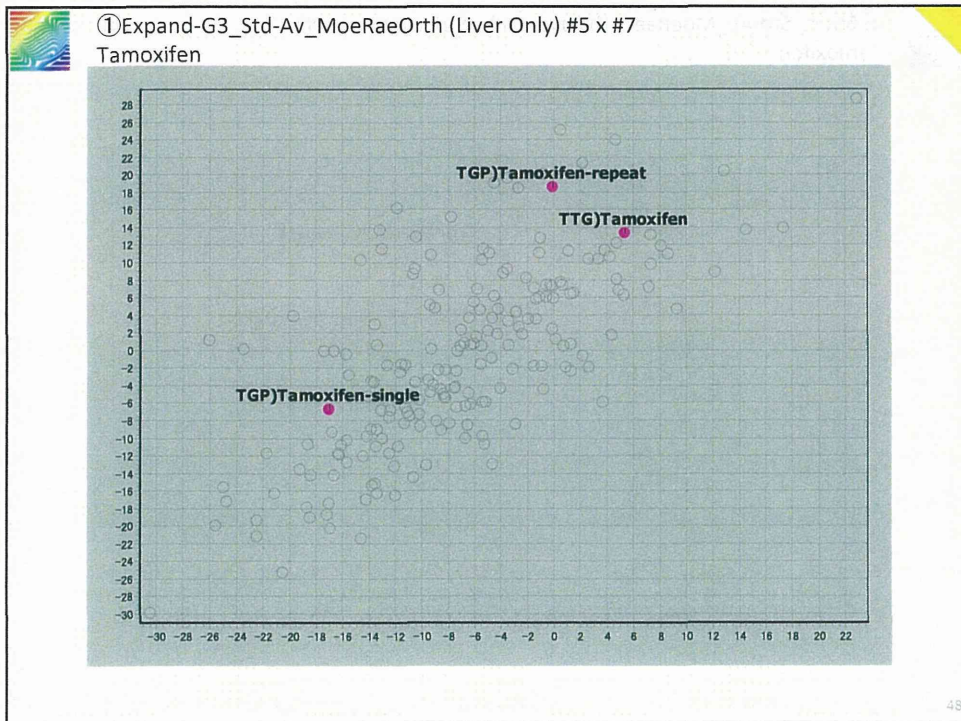
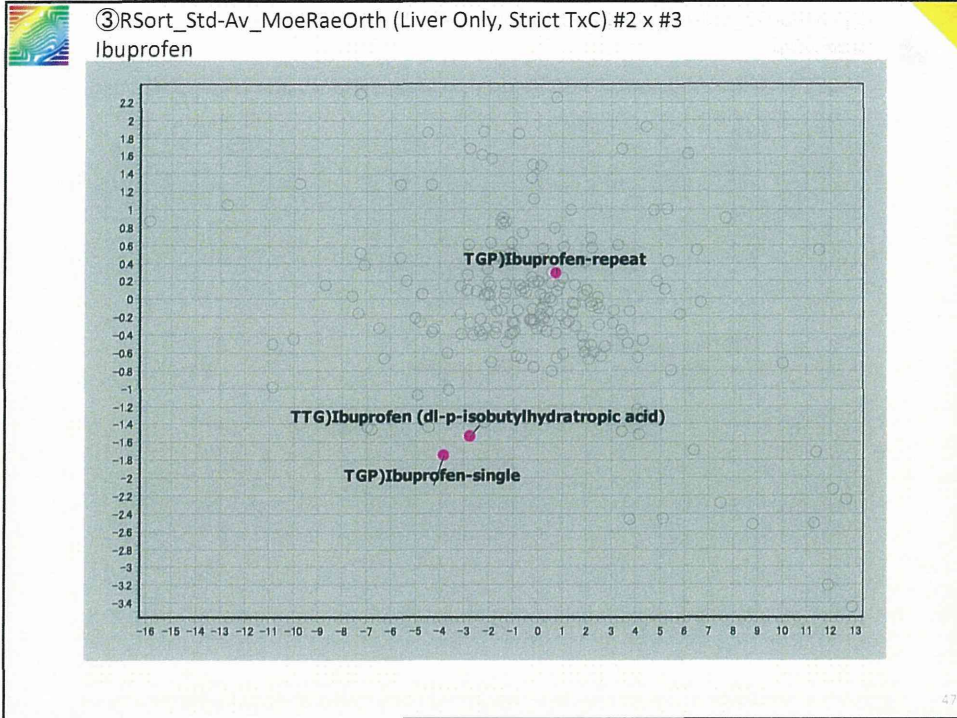


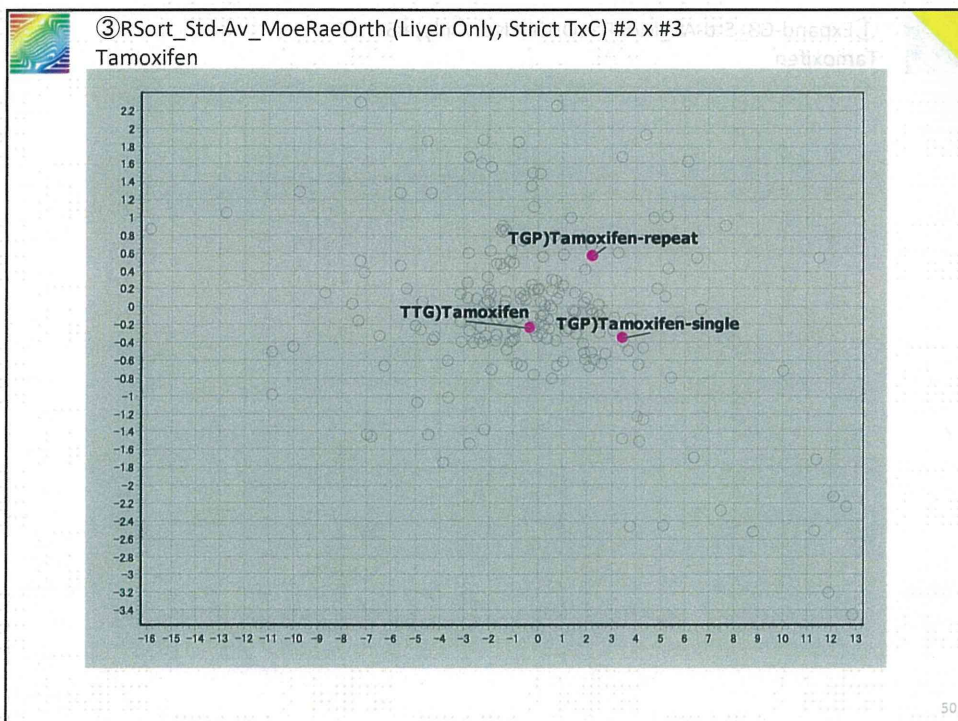
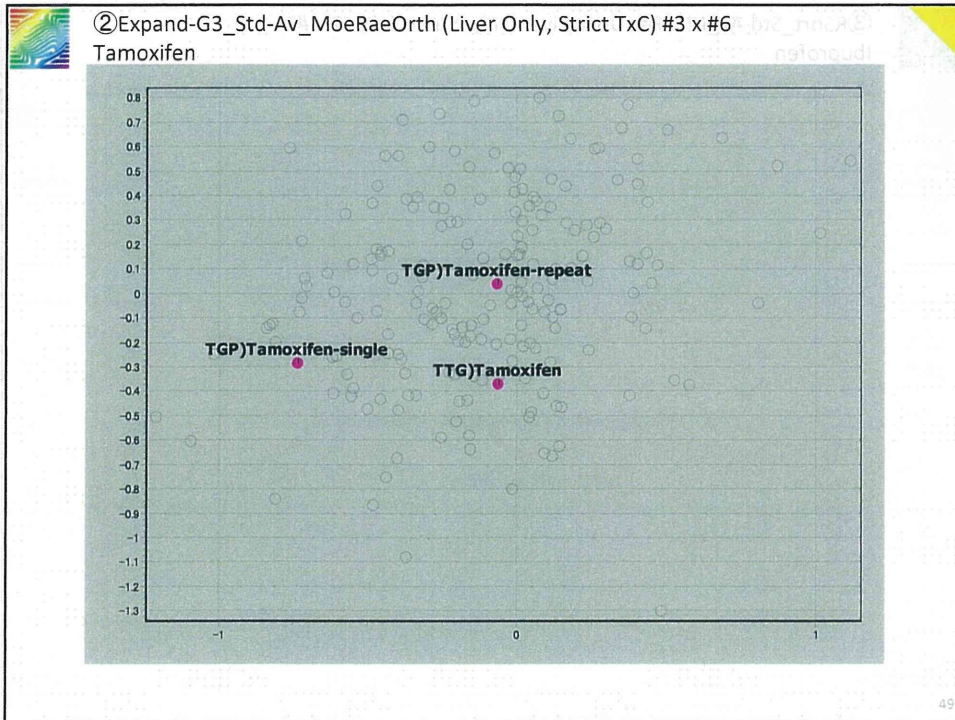
③RSort\_Std-Av\_MoeRaeOrth (Liver Only, Strict TxC) #2 x #3  
Acetaminophen



44



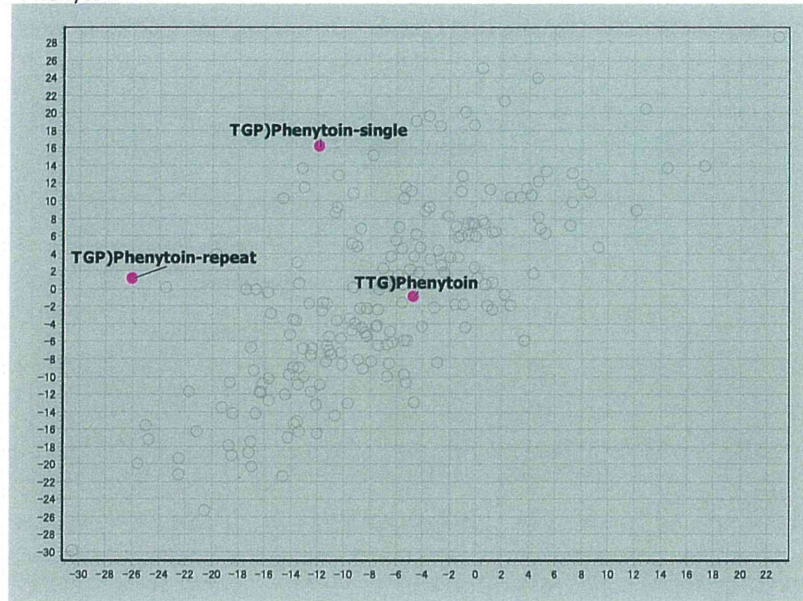








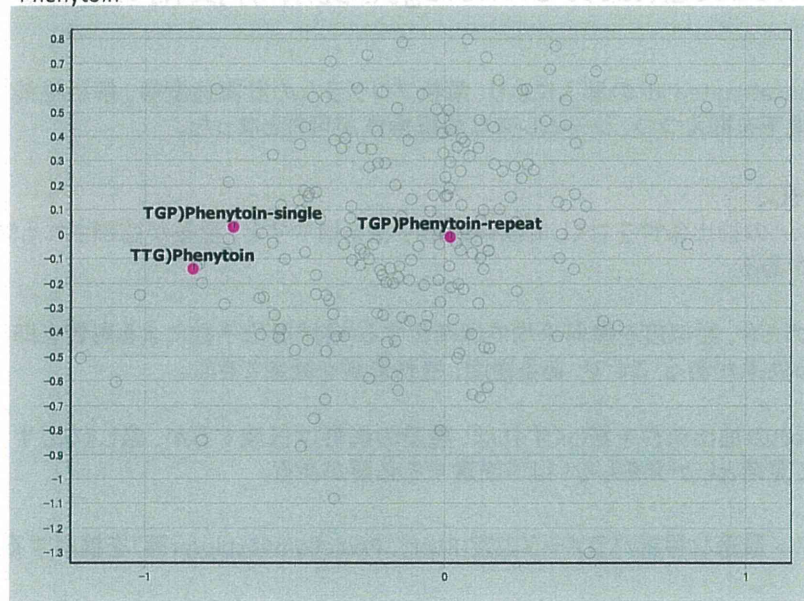
①Expand-G3\_Std-Av\_MoeRaeOrth (Liver Only) #5 x #7  
Phenytoin



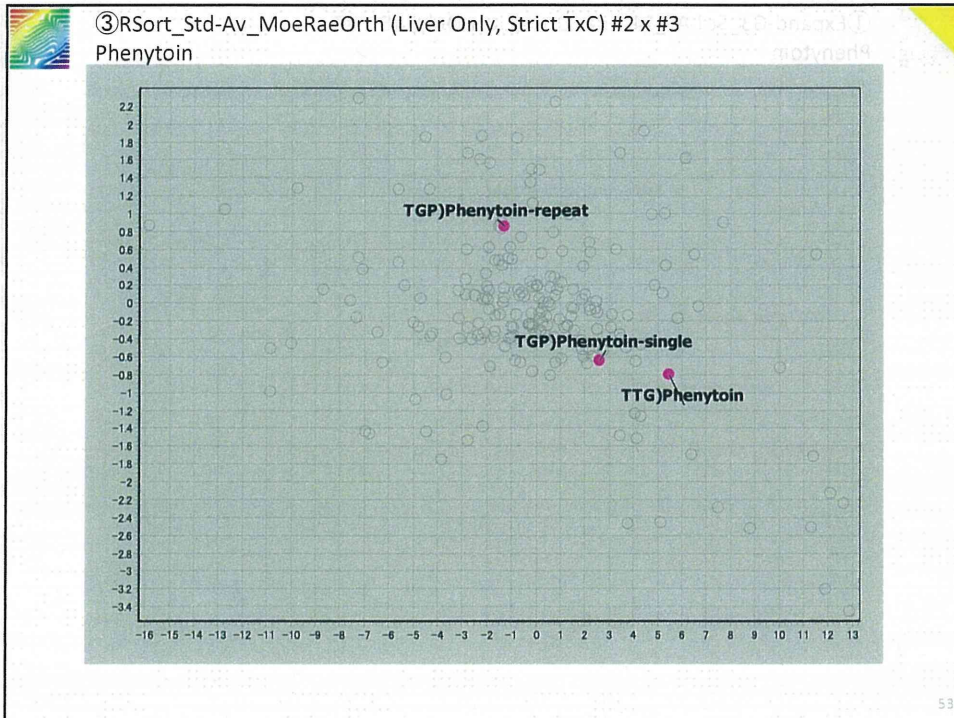
51



②Expand-G3\_Std-Av\_MoeRaeOrth (Liver Only, Strict Tx) #3 x #6  
Phenytoin



52



## MouseとRatのOrtholog発現解析技術の開発

“Concatenated Id”の導入により、解析プログラムの計算速度等、解析性能の低下を抑えつつ、シームレスな解析操作が可能となった。

ただし、

- RSortの抽出条件を甘くして候補遺伝子数を増やすと、種差が強調される傾向がある。
- このとき、類似度を幾何平均で標準化する(候補遺伝子数による影響を抑える効果がある)ことで、ある程度、種差影響を軽減できる。
- RSortの抽出条件を厳しくすれば、種差の影響は低減するが、厳しくなりすぎて見落としが発生しないよう注意する必要がある。

>>>> 最適な解析パラメータ設定(RSort、PercellomeExplorer等)を検討する。

54



## H24年度の成果

- 国際的なソフトウェア共通基盤Garudaへの対応  
～REST API of PercellomeWeb service ver.1
- 一般的な(外部Spikeが追加されていない)測定済み  
マイクロアレイデータの絶対量推定技術の開発
- TGPラット・トキシコゲノミクスデータの統合  
～シームレスなOrtholog発現解析技術の開発

55



## H25年計画

- 国際的なソフトウェア共通基盤Garudaへの対応  
～REST API of PercellomeWeb service ver.2の開発・  
リリース
- TGPラット・トキシコゲノミクスデータの完全統合と、  
Ortholog発現解析へ対応させるための解析ソフト  
ウェアアップデート(SnCalcのラット適用を含む)
- 各化学物質による生体反応の特異的要素のハイス  
ループット抽出・解析技術の開発  
～PercellomeExplorerの拡張

56

厚生労働科学研究費補助金（化学物質リスク研究事業）

分担研究報告書

化学物質の有害性評価手法の迅速化、高度化に関する研究  
—網羅的定量的大規模トキシコゲノミクスデータベースの維持・拡充と  
毒性予測評価システムの実用化の為のインフォマティクス技術開発—

分担研究課題：「システムトキシコロジー解析基盤の研究開発」

研究分担者 北野 宏明

特定非営利活動法人システム・バイオロジー研究機構 会長

研究要旨

本研究では、システムバイオロジーをトキシコロジーに応用し、大規模かつ網羅的なデータを有効に利用し、化学物質の毒性の機序の理解を推進すると同時に、遺伝子発現データに反映される生理学的影響から、その原因物質が、遺伝子制御ネットワークのどの部分に影響を及ぼしているのかなどを同定する手法を開発する。

平成 24 年度は、遺伝子制御関係推定アルゴリズムとして、複数のアルゴリズムの併用を行う方式を開発し、現状での最も優れた方式と同等の結果を安定的に維持することに成功したほか、クラスタリングソフトウェア (AGCT) の Percellome データへの最適化や計算規模及び速度の向上を実施した。また Percellome データベースの、国際的なソフトウェアの共通基盤 Garuda Platform 上への実装を進め、国際的な標準化や普及のための準備を行った。

A. 研究目的

本研究では、大規模かつ網羅的なデータを有効に利用し、化学物質の毒性の機序の理解を推進すると同時に、遺伝子発現データに反映される生理学的影響から、その原因物質が、遺伝子制御ネットワークのどの部分に影響を及ぼしているのかなどを同定する手法を開発する。この手法によって、大量の文献情報などから集約したボトムアップな知識の集積によって毒性の機序が理解できると同時に、大規模データから毒性に関連して変動する遺伝子群の同定とそれらの相互作用関係、さらには、各々の生理学的影響または毒性に相関すると思われる遺伝子の変動に対して、それらの上流に存在するどの遺伝子群やどの遺伝子群を制御

する転写因子や受容体が重要であるかを判別することが出来る。

同時に、これら一連の手法は、ソフトウェアとして実現し、広く利用されることが望ましい。そこで、我々が、提唱し、開発が進んでいる、Garuda Common Platform に準拠した一連のソフトウェアツールを開発する。

B. 研究方法

本研究は、システムバイオロジーをトキシコロジーに応用する研究であり、従来の手法ではとらえきれなかったネットワークを基盤とした毒性の理解を可能とする。このため、大規模データを利用する手法と大規模文献情報などをいかに集約するかに関

する研究、そしてそれらを統合的なプラットフォームに実装し広く利用できるようにする研究によって構成される。よって、この研究は、以下の項目に区分できる：

- (1) 大規模データからの遺伝子制御関係推定アルゴリズムの開発と改良
- (2) 大規模データからの状態制御遺伝子群推定アルゴリズムの開発と改良
- (3) 大規模文献情報などの集約手法と伝達手法の研究
- (4) **Garuda Platform** とその関連ソフトウェアの研究開発

#### 1: 大規模データからの遺伝子制御関係推定アルゴリズムの開発と改良

**Percellome** や多くの研究で利用されている遺伝子発現データを利用し、遺伝子間の発現変動の相互関係と因果関係ネットワークの構築を可能とする手法の研究を行う。現在、遺伝子発現データからのネットワーク抽出の研究は多く行われているが、その精度は高くない。例えば、広く使われているベイズ統計を利用した方法は、比較的少数の遺伝子に対して大量のデータが存在するときには適切な方法とは言えない。このような状況では、相互情報量 (**Mutual Information**) を利用する手法の方が適しているが、その計算の具体的な手法により、推定結果の精度などは大きく違ってくる。我々では、既に、複数の計算手法の組み合わせを利用して、これらの現存手法より、精度の高い手法を開発した。本研究では、この手法をさらに改良し、より精度の高いものに改良する。同時に、発現の位相差、時間差、ゲノム情報などを統合し、相関関係ではなく因果関係に関するネットワークを高精度で構築可能な手法の開発を行う。

#### 2: 大規模データからの状態制御遺伝子群推定アルゴリズムの開発と改良

システム毒性学で重要なことは、ネットワークとして記述された生体システムが、毒性物質の影響でどのように「制御」されてしまうかを理解すること (フォワードアナリシス) であり、また、ネットワークを特定の状態に遷移させた原因は、どの遺伝子群への影響によるものであるかを特定できるということ (バックワードアナリシス) である。フォワードアナリシスは、研究項目 1 と 3 で実現する。この研究項目では、ネットワークを制御することができる遺伝子群の同定を大規模データから行う手法の研究を行う。

我々は、すでにプリオン病のデータを利用して、遺伝子相互作用ネットワークの分析から、プリオン病と同等の遺伝子発現変化を生じさせる少数のドライバー遺伝子を同定し、その効果をシミュレーションで確認している (図 1)。同時に、発病した状態に対応するネットワークを、正常な状態にもどす制御を可能とする別の数個の遺伝子を同定する手法も開発している。現在のところ、この手法は研究の初期段階であり、その計算手法や予測精度、応用範囲など、多くの研究の余地が残されている。本研究では、この手法に関する研究を進めることで、システム毒性研究における重要な手法とすることを目指す。この研究には、遺伝子制御ネットワークの変化だけではなく、影響を及ぼすドライバー遺伝子の同定とそれに関わる転写因子や受容体の同定なども含める。また、SNPs などの違いがこのネットワークの挙動に影響を及ぼすことが分かった場合には、これらの影響を組み込む手法の開発も試みる。

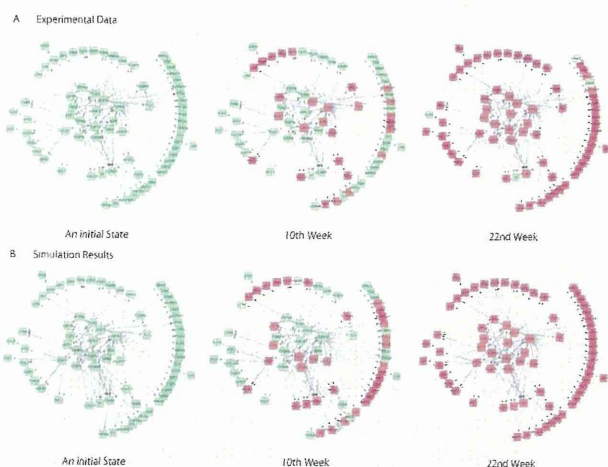


図1：(A) プリオン病の進行過程での遺伝子発現変化の実測値をネットワーク上に表示した。(B) 我々の手法で、同定した3個の遺伝子の過剰発現からの遺伝子発現の変化をシミュレーション結果。

### 3: 大規模文献情報などの集約手法と伝達手法の研究

フォワードアナリシスでは、大規模データのみならず、大量の文献やデータベースなどを総動員して精度の高い分子相互作用ネットワークと遺伝子制御ネットワークを構築する必要がある。すでに我々は、これらに関するソフトウェアである CellDesigner や Payao などを構築して、このプロセスの効率化と高精度モデルの構築ノウハウを蓄えてきた。この蓄積から、次のステップとして大量に存在する文献などの情報を、知的に集約し、モデル構築を行う研究手法などを提示し、可能な範囲で自動的にモデルを構築する情報を抽出する手法の開発が必要であると結論している。この研究項目は、このプロセスに関する研究とそのコンセプトを実証するソフトウェアの試作を行っている。

### 4: Garuda Platform に準拠する毒性解析ソフトウェアの開発

本研究で開発される手法を実装するソフトウェアや Percellome などは、国内外の幅広い研究者に利用されることで、トキシコロジーに関連する研究の発展とその成果の利用を促進する必要がある。我々は、国際的なソフトウェアの共通基盤として Garuda Platform を開発している。ここでは、この Garuda Platform をトキシコロジーの分野へと応用することを可能とする一連の開発と普及・啓蒙を行う。

### C. 研究成果

遺伝子制御関係推定アルゴリズムでは、複数のアルゴリズムの併用を行う方式を開発し、現状での最も優れた方式と同等の結果を安定的に維持することに成功している。さらに、より一層の大規模化に備え、この性能を維持しつつ、情報エントロピーを利用しながら計算量を減少させる手法の開発に成功した。

状態制御遺伝子群推定アルゴリズムに関しては、以前から開発している A Gene geometric Clustering Tool (AGCT) で Percellome データの特徴を考慮に入れた解析を容易に実行できるようにするための開発を行った。具体的には、全データからのドーズ選択(細胞選択)、時計遺伝子効果削除、スペクトラルクラスタリングや主成分分析の表示の上で6つのアルゴリズムによるクラスタリング等の機能を追加した。これらの機能では Percellome データのニーズにそくし、事前データ削除を行わずに全データを計算した後、必要な部分だけをユーザが取り出せるようになっている。更に大規模データをより短時間で計算できるようにした。

制御遺伝子群の同定についての予備検討では、幾つかの事例において、実験的に確

認められている遺伝子群と合致することなどを確認した。

さらに、Garuda Platform の開発者向けパッケージの配布開始（2013年2月）にあわせ、ネットワークの推定やAGCTなど一連のアプリケーションをGarudaに準拠し、毒性学に利用できるように対応したほか、オルソログ制御領域分析ソフトウェアSHOE（Sequence Homology in Higher Eukaryotes）に関してはオンラインでGaruda経由でサービス提供できるように準備を進めた。またPercellome DBのGaruda上への実装に協力し、Percellome DBの国際的普及への準備を行った。

#### D. 考察

毒性学にシステムバイオロジーを適用する一連の手法が具体的な形となっており、複数のソフトウェアの連動が実現されつつある。

Garudaは、今後極めて重要なプロジェクトに成長すると思われる。ソフトウェアを開発する側の研究者のみならず、ユーザとなる製薬企業や生命科学の基礎研究に携わるものにとっても、重要なPlatformであるとの認識が広まっている。このGaruda Platformにシステム毒性学の研究を可能とするソフトウェア群を実現する事は、今後の研究の展開にとって非常に重要であると考える。

#### E. 結論

本研究は、システムバイオロジーを毒性学に応用しようとするものであり、そのために一連の方法論とソフトウェアを開発し、応用するものである。現在、この目標に対して、順調に展開している。すでに幾つかのソフトウェアが、Garuda Platform上に実装され連動するなどの成果をあげている。

今後、これらのソフトウェアを利用した具体的適用例を作る事や普及などの展開にも力を入れたい。

#### F. 研究発表

##### (1) 論文発表

1. Martin H. Schaefer; Tiago J. S. Lopes, Nancy Mah, Jason E. Shoemaker, Yukiko Matsuoka, Jean-Fred Fontaine, Caroline Louis-Jeune, Amie J. Einfeld, Gabriele Neumann, Carol Perez-Iratxeta, Yoshihiro Kawaoka, Hiroaki Kitano, Miguel A. Andrade-Navarro. Adding Protein Context to the Human Protein-Protein Interaction Network to Reveal Meaningful Interactions. *PLOS Computational Biology*. 9, 1, 2013.
2. Crespo, I.; Roomp, K.; Jurkowski, W.; Kitano, H.; del Sol, A. Gene regulatory network analysis supports inflammation as a key neurodegeneration process in prion disease. *BMC Systems Biology*, 6:132, 2012.
3. Shoemaker, J.; Fukuyama, S.; Einfeld, A. J.; Muramoto, Y.; Watanabe, S.; Watanabe, T.; Matsuoka, Y.; Kitano, H.; Kawaoka, Y. Integrated network analysis reveals a novel role for the cell cycle in 2009 pandemic influenza virus-induced inflammation in macaque lungs. *BMC Systems Biology*, 6:117, 2012.
4. Shoemaker, J. E., Lopes, T. J., Ghosh, S., Matsuoka, Y., Kawaoka, Y., Kitano, H. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics*, 13 (1): 460, 2012.
5. Martijn P. van Iersel, Alice C. Villeger, Tobias Czauderna, Sarah E. Boyd, Frank T. Bergmann, Augustin Luna, Emek Demir, Anatoly Sorokin, Ugur Dogrusoz, Yukiko Matsuoka, Akira Funahashi, Mirit I. Aladjem, Huaiyu Mi, Stuart L. Moodie, Hiroaki Kitano Nicolas Le Novere, and Falk Schreiber. Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinformatics*. 28 (15): 2016-21, 2012. doi: 10.1093/bioinformatics/bts270,

published online May 10, 2012.

6. Carl-Fredrik Tiger, Falko Krause, Gunnar Cedersund, Robert Palmér, Edda Klipp, Stefan Hohmann, Hiroaki Kitano and Marcus Krantz. A framework for mapping, visualisation and automatic model creation of signal-transduction networks. *Molecular Systems Biology*. 8: 578, April 24, 2012.

## (2) 学会発表

1. Natalia Polouliakh and Hiroaki Kitano “Discovery of gene network by clustering and promoter analysis. *The Journal of Toxicological Sciences* ISNB:0388-1350 Vol:37 (Supplement 1) S31. The Japanese Society of Toxicology.
2. Kitano, H. Will engineering play the lead role in drug discovery in 2030? BioMelbourne breakfast, Cinema 1, Melbourne, Australia, June 5. (invited)
3. Kitano, H. Software Platform for Systems Drug Discovery. Bio-IT World Asia Conference 2012, Marina Bay Sands, Singapore, June 8, 2012. (invited)
4. Kitano, H. Systems Drug Design and Garuda Software Platform. Network Biology SIG: On the Analysis and Visualization of Network in Biology (NetBio SIG), ISMB 2012, Long Beach Convention Center, USA, July 13, 2012. (invited)
5. Kitano, H. HD-Physiology, Garuda, and Computational drug side-effects prediction. Talk at FDA, FDA, Silver Spring, USA, July 17, 2012. (invited)
6. Kitano, H. Biological Robustness. Seminar at University of Toronto, University of Toronto, Canada, Aug. 20, 2012. (invited)
7. Kitano, H. Systems Biology powered by Artificial Intelligence. PRICAI-2012: 12th Pacific Rim International Conference on Artificial Intelligence (via skype, invited), Pullman Hotel, Kuching, Malaysia, Sep. 7, 2012. (invited)

8. Kitano, H. VPH in industrial research. VPH 2012, Savoy Place, London, UK, Sep. 20, 2012. (invited)
9. 北野宏明. システム創薬とGarudaプラットフォームの概要. BioJapan 2012, パシフィコ横浜, Oct. 11, 2012, (invited)
10. Kitano, H. Systems biomedicine and their computational platforms. FOSBE 2012, Institute for Advanced Biosciences, Keio University, Oct. 21, 2012. (invited keynote)
11. 北野宏明. システムバイオロジー概論. 第6回 KAST システムバイオロジー講座, かながわサイエンスパーク, Nov. 2, 2012. (invited)
12. Kitano, H. “Systems Toxicology”, An invited Talk at DSTO, Defence Science and Technology Organisation (DSTO), Department of Defence, Australian Government, Melbourne, Australia, Dec. 4, 2012. (invited)
13. 北野宏明. Data-Driven Network-Based Biomarker Discovery. 第35回 日本分子生物学会年会 バイオテクノロジーセミナー, 福岡国際会議場, Dec. 12, 2012. (invited)

## G. 知的財産権の出願・登録状況

1. 特許取得  
なし
2. 実用新案登録  
なし
3. その他  
なし



班会議平成25年2月8日

クラスタリングとプロモータ解析によ  
りネットワーク検出のアプローチ  
Gene Clustering and Promoter  
Analysis

Sony Computer Science Laboratories Inc.

Natalia Polouliakh and Hiroaki Kitano

2013.02.08

## Agenda 1

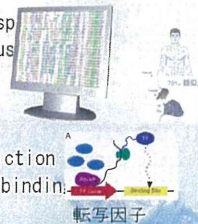
- ◆ **AGCT** A Gene *geometric* Clustering Tool
  - ◆ Clustering Percellome data based on similarity of gene expression profile. Application on TCDD and TCDF chemicals.  
(2, 3, 7, 8-Tetrachlorodibenzo-p-dioxin and 2, 3, 7, 8-Tetrafurran)
    - ◆ Data loading and normalization
    - ◆ Coding signal to slope and wavelet
    - ◆ Computer similarity matrix
    - ◆ Processing spectral manifold and PCA variance
    - ◆ Clustering data with 6 algorithms:
      - ◆ Bregman (Iterative) K-means (optimizing K)
      - ◆ Affinity Propagation
      - ◆ Expectation Maximization
      - ◆ Hierarchical clustering
      - ◆ Complete Positive Factorization
      - ◆ Non-negative matrix factorization
  - ◆ AGCT-Garuda project connection
  - ◆ Publication: Natalia Polouliakh, Richard Nock, Frank Nielsen, Hiroaki Kitano "G-Protein Coupled Receptor Signaling Architecture of Mammalian Immune" CellsVol.4No.1pages e4189PLoS ONEJanuary2009.

## Agenda 2

### ◆ SHOE Sequence Homology in Higher Eukaryotes

#### ◆ Phylogenetic footprinting on human-mouse-rat for discovery of transcription regulation network.

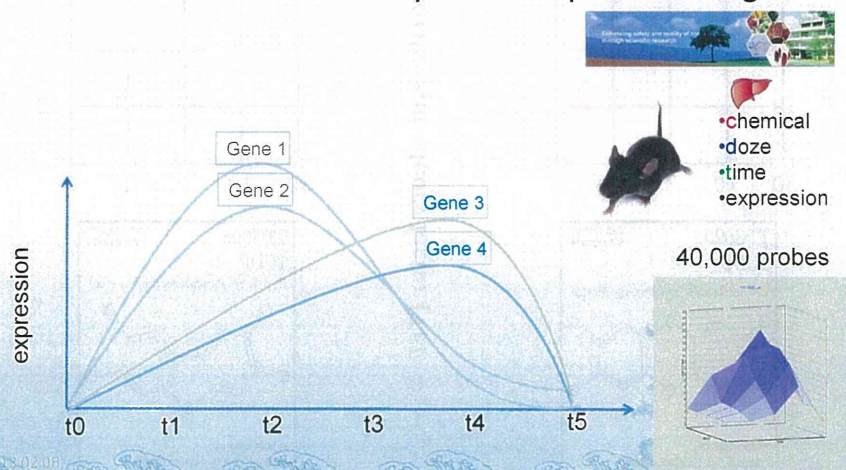
- Extract orthologous regulatory regions of three species
- Perform pair-wise alignment between human and mouse and between human and rat.
- Perform multiple alignment on conserved regions over three species.
- Perform transcription factor binding motifs prediction using publicly available transcription factor binding matrices



### AGCT A Geometric Clustering Tool

## Atlas of Cell Life by AGCT

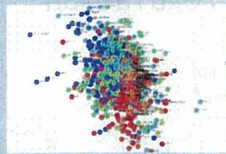
- ◆ AGCT reconstructs gene network basing on the similarities of the expression profiles of genes



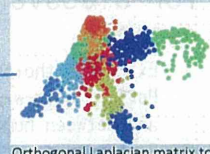
## Processing data on AGCT

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\*N matrix



Spectral clustering:  
M\*M matrix



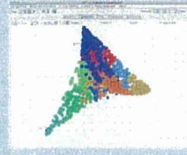
Orthogonal Laplacian matrix to compute one dimension per cluster/gene

### Examples of different network topologies

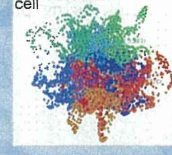
Mouse Stem cell



TCDD affected mouse liver cell

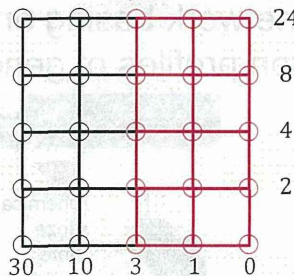


Influenza affected mouse bronchi cell

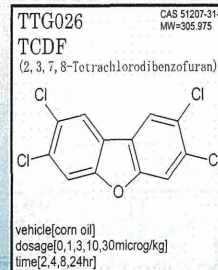
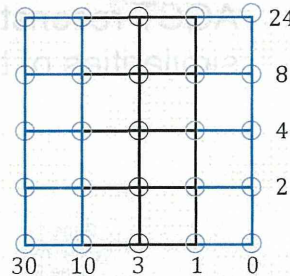


2,3,7,8-Tetrachlorodibenzo-p-dioxin and 2,3,7,8-Tetrafuran chemicals are known to cause birth defects, immunotoxicity and cancer through the activation of Ahr pathway

TCDD



TCDF



Toxicity equivalency factor (TEF 0..1)

## Percellome Dataset

- ◆ 全発現プロファイルが20,000プローブを含む
  - ◆ 初期反応遺伝子 (responsive transcription) (~27%)
  - ◆ Circadian 遺伝子 (basal transcription)
  - ◆ 無反応遺伝子 (not circadian)
  - ◆ 実験誤差

Needle in the haystack problem



## AGCT analyses data without constraints. **Suitable to Percellome.**

- ◆ Automatic selection of dose to analyze from the entire dataset.
- ◆ Circadian effect subtraction (J. Kanno)
- ◆ Large data size processing  $40,000 \times (2 \dots n)$
- ◆ Unsupervised clustering
- ◆ Filtering analyzed clusters after processing