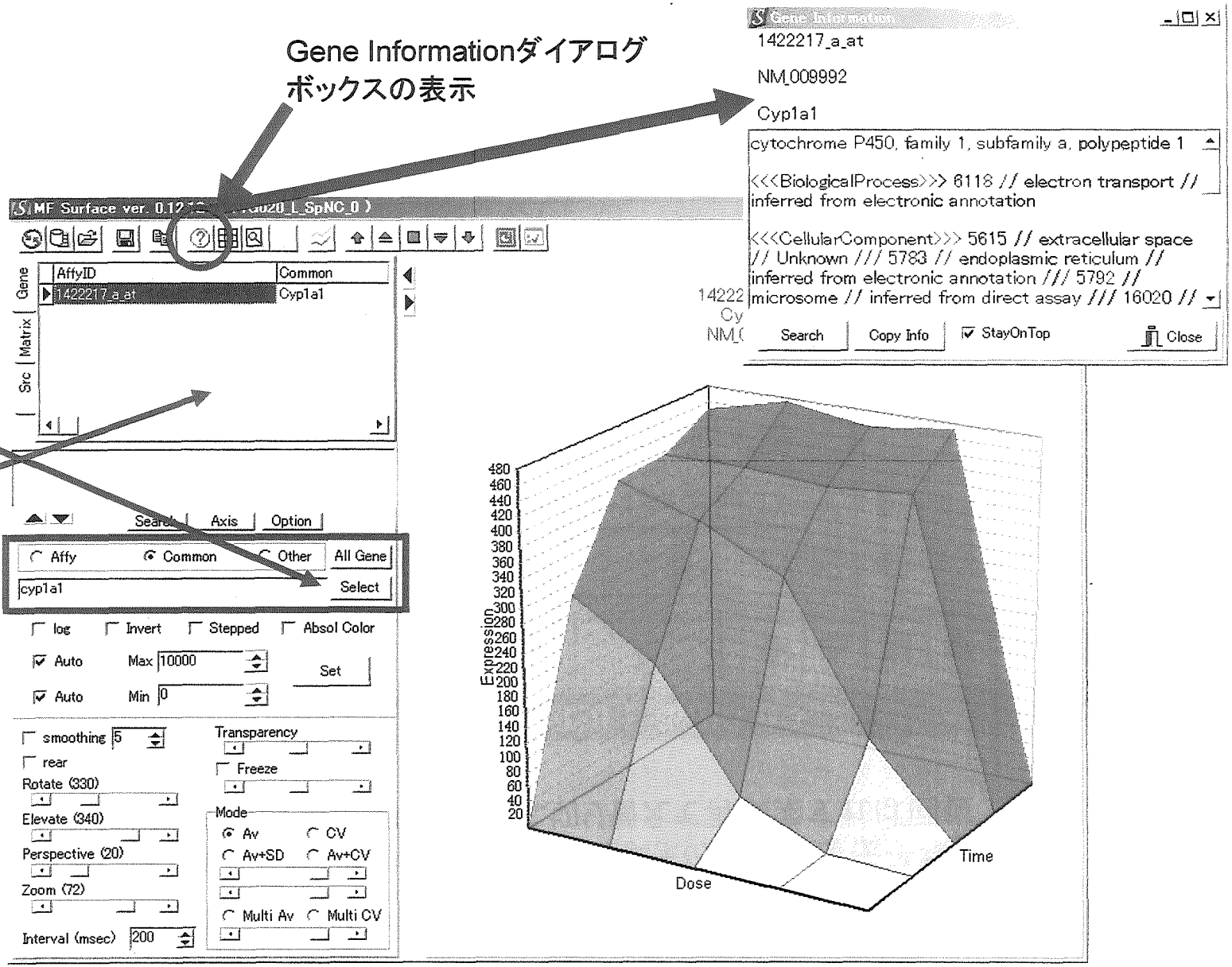
遺伝子の検索、Annotation情報の表示

検索対象のフィールド(ID[AffyID], 遺伝子名[common]、全文検索[Other])を指定して、キーワードを入力の上、Selectボタンを押す。

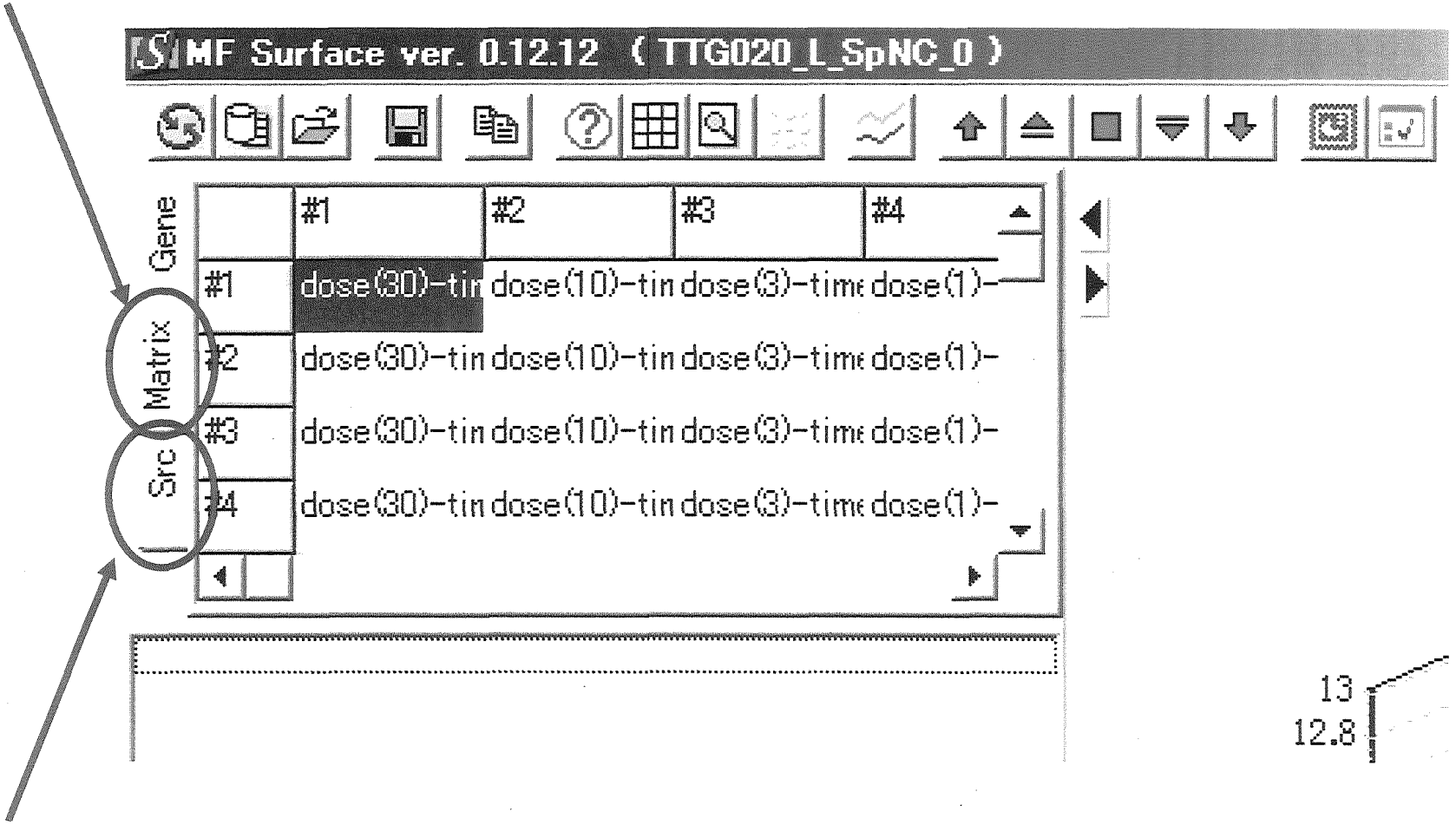
結果は上のGridに表示される。

Gene Informationダイアログボックスの表示



データおよびデータ配列の確認

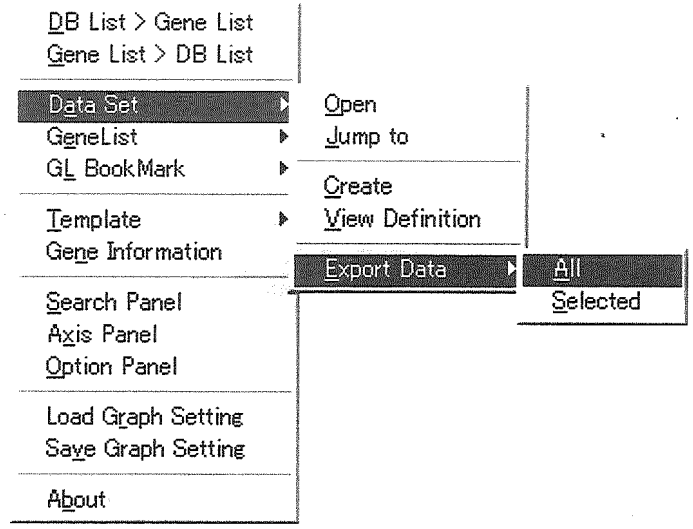
データ配列を確認するには、Matrixタブに切り替える。（このままの状態で遺伝子検索などのを行うとエラーがでることがあるが、Geneタブに切り替えて再検索すれば回復する）

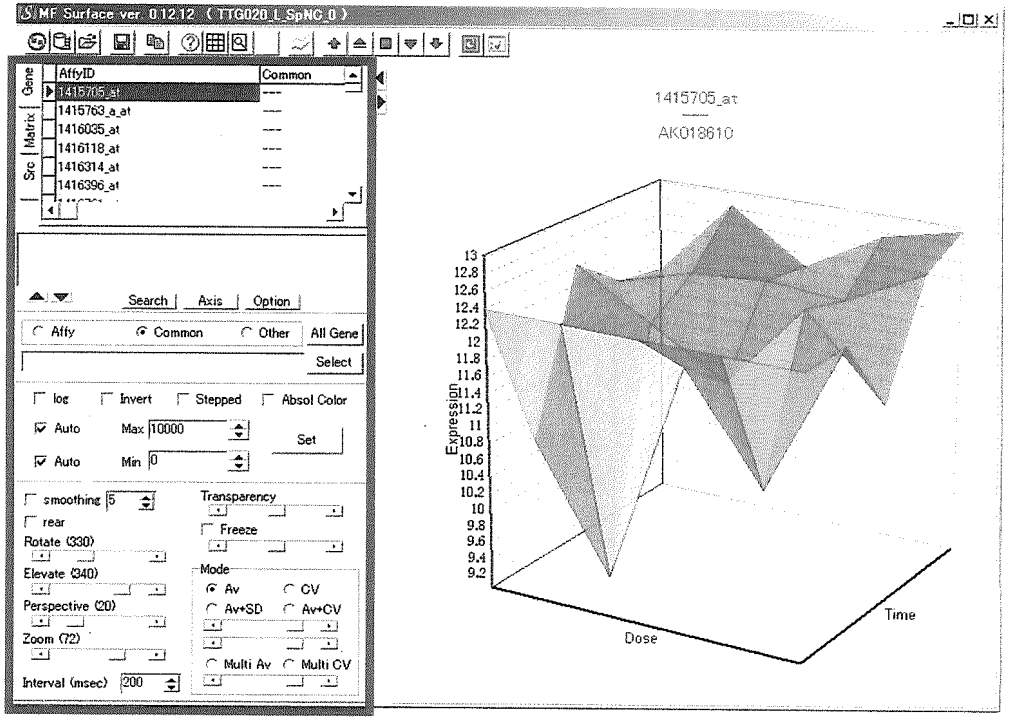


平均データを確認するには、Srcタブに切り替える。（このままの状態で遺伝子検索などのを行うとエラーがでることがあるが、Geneタブに切り替えて再検索すれば回復する）

データのエクスポート

左側の操作パネル部分で右クリックし、コンテキストメニューから「DataSet」→「Export Data」→「All」もしくは「Select」を選択する。
「All」では、全てのデータを出力し、「Select」では読み込み済みの遺伝子リストにあるデータのみ出力する。

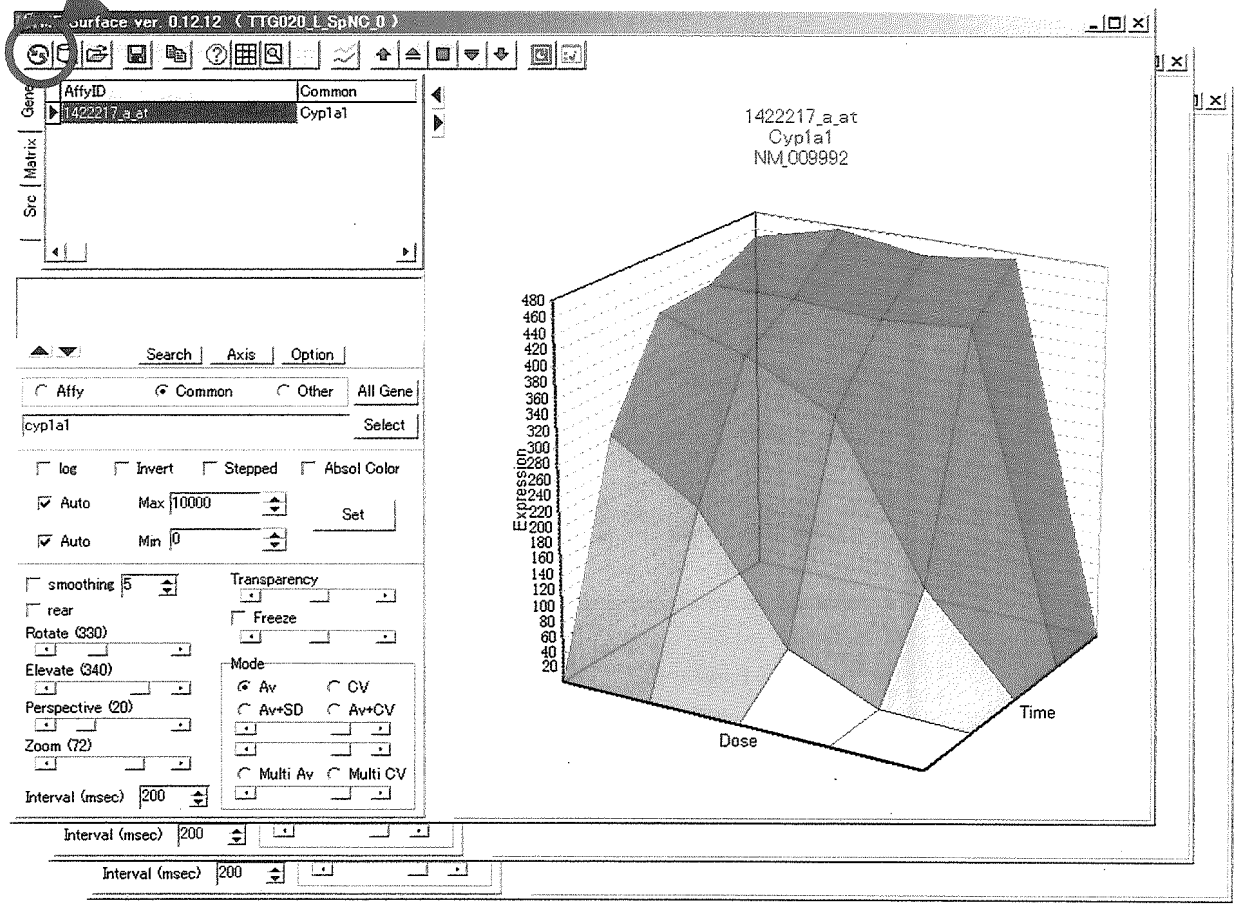




複数のMFSurfaceの同期操作

複数のMFSurfaceを立ち上げ、別々のデータセットを読み込んだ状態で、選択した遺伝子の情報やグラフ設定などを同期することが出来る。ただし、複数のデータセット間比較は後に開発したMSVを用いた方が便利

同期させたい全てのMFSurfaceにおいてリンクボタンを押し下げておく
同期させる内容はリンクボタン右クリックで設定可能



国立医薬品食品衛生研究所　毒性部　御中

NTT COMWARE

遺伝子発現パターン
解析結果報告書

2004年3月31日（水）

NTTコムウェア株式会社
ビジネス創出部

目次

NTT COMWARE

1. 期待値最大化法を利用したクラスタリング
2. 次元リダクション
3. TMF計算
4. 密度ベースと階層的クラスタリングの融合
5. 関数近似手法

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（１）Uteroデータに対するクラスタリング分析

入力データ

実験条件（時間／ｄｏｓｅ）毎にｎ匹（２－４匹）のマウスによる発現量を単純平均した値

- Ｇｅｎｅ数　：　１２，４８８個
- 変数　　：　４（時間毎）Ｘ４（ｄｏｓｅ毎）＋１（０値）＝１７変数

クラスタリング手法

期待値最大化法（ＥＭ法）　※ＥＭ＝Expectation Maximization

マハラノビス距離を使用し、統計的手法によりノイズの影響を受けにくい
結果から求められる統計量（ＡＩＣ：赤池情報量規準）により、最適なクラスタ数を求められる

データ加工手法

（１）加工なしによるクラスタリング
（２）発現量の対数変換によるクラスタリング

1．期待値最大化法を利用したクラスタリング

NTT COMWARE

（1）Uteroデータに対するクラスタリング分析

分析結果　・・・　（1）加工なしによるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	24	-101.6577	881.32
3	37	-91.8425	1199.69
4	43	-86.8586	1527.72
5	50	-83.5964	1859.19

ＡＩＣ (Akaike Information Criteria)

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

収束ステップ数より、クラスタの数を増やしていくと収束しにくい結果が得られた。
これは、発現量が一様に分散しており、カテゴリ分けしにくい傾向があると推測される。

また、ＡＩＣの値から　クラスタ数＝２が最適　という、

　発現する　／　発現しない

という２つのクラスタに分かれたと思われる　現実的な解ではない結果　になった

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（１）Uteroデータに対するクラスタリング分析

分析結果　・・・　（２）発現量の対数変換によるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	12	-4.1891	686.38
3	5	0.0884	1015.82
4	3	1.0533	1351.89

ＡＩＣ (Akaike Information Criteria)

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

（１）の結果を踏まえて、データを対数変換することにより、発現量を圧縮したあとにクラスタリングを実施した。収束ステップ数をみるに、その効果はあったと思われる。

しかしながら、ＡＩＣの値から　クラスタ数＝２が最適　という、

　発現する　／　発現しない

という２つのクラスタに分かれたと思われる　現実的な解ではない結果　になった

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（１）Uteroデータに対するクラスタリング分析

結果の考察

２種類の方法により、クラスタリングを実施してきたが、両方とも現実的な解でないという結果に終わった。

これらの原因として、

- 変数が多すぎることによる影響
- ノイズによる影響

などが考えられる。

これらのことから、これらの影響を除去するために、次のような方法が有効ではないのかと考えた

（１）変数が多すぎることによる影響を避けるために、何らかの方法で変数を減らす

（２）ノイズによる影響を避けるために、Ｐｒｅｓｅｎｃｅが多くみられるＧｅｎｅに絞ってクラスタリングを実施する

今回は（１）について、その方法についてのアプローチを検討した

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（２）ＴＴＧ８データに対するクラスタリング分析

分析結果　・・・　（１）加工なしによるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	17	-90.4904	786.98
3	39	-82.1528	1072.31
4	50	-77.3473	1364.69

ＡＩＣ (Akaike Information Criteria)

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

収束ステップ数より、クラスタの数を増やしていくと収束しにくい結果が得られた。
これは、発現量が一様に分散しており、カテゴリ分けしにくい傾向があると推測される。

また、ＡＩＣの値から　クラスタ数＝２が最適　という、

　発現する　／　発現しない

という２つのクラスタに分かれたと思われる　現実的な解ではない結果　になった

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（２）ＴＴＧ８データに対するクラスタリング分析

分析結果　・・・　（２）発現量の対数変換によるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	6	-9.7482	625.50
3	12	-4.4372	916.87
4	14	-1.2714	1212.54

ＡＩＣ (Akaike Information Criteria)

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

収束ステップ数より、クラスタの数を増やしていくと収束しにくい結果が得られた。
これは、発現量が一様に分散しており、カテゴリ分けしにくい傾向があると推測される。

また、ＡＩＣの値から　クラスタ数＝２が最適　という、

　発現する　／　発現しない

という２つのクラスタに分かれたと思われる　現実的な解ではない結果　になった

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（３）ＴＴＧ９データに対するクラスタリング分析

分析結果　・・・　（１）加工なしによるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	4	0.0129	605.97
3	4	0.1445	907.71
4	4	0.081	1209.84

※クラスタ数を幾つに指定しても「クラスタリング停止：最小分数条件が１つ以上のクラスター内にあります。」発生した。単一ポイントがクラスタとみなされたと考えられる。

分析結果　・・・　（２）発現量の対数変換によるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	3	−1.4189	608.84
3	3	−1.4189	910.84
4	3	−1.4188	1212.84

ＡＩＣ（Akaike Information Criteria）

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

１．期待値最大化法を利用したクラスタリング

NTT COMWARE

（４）ＴＴＧ１０データに対するクラスタリング分析

分析結果　・・・　（１）加工なしによるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	5	0.0026	605.99
3	4	0.0505	907.90
4	4	0.0698	1209.86

※クラスタ数を幾つに指定しても「クラスタリング停止：最小分数条件が１つ以上のクラスター内にあります。」発生した。単一ポイントがクラスタとみなされたと考えられる。

分析結果　・・・　（２）発現量の対数変換によるクラスタリング

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	3	-1.4189	608.84
3	3	-1.4189	910.84
4	3	-1.4188	1212.84

ＡＩＣ (Akaike Information Criteria)

与えられたデータの範囲でより近い数式モデルを作るための統計指標。小さいほどよい

２．次元リダクション

NTT COMWARE

（１）主成分分析の因子によるリダクション

ＥＭ法をそのまま適用すると収束しにくく、発現量を圧縮しても

発現する　／　発現しない

という２つのクラスタに分かれたことから、主成分分析を実施し、何個の因子で１７変数が説明できるかを固有値をみることにより考察した

因子	固有値	累積寄与率
第1因子	15.75	92.6%
第2因子	0.73	96.9%
第3因子	0.21	98.2%
第4因子	0.07	98.6%
・	・	・
第17因子	0.01	100.0%

固有値：データのばらつきの割合

累積寄与率：その固有値までで説明できるばらつきの割合

上記の　第１因子の累積寄与率＝９２．６％　より、１７変数から導かれる第１因子でデータのほとんどを説明できてしまうということが言えることから、単純な主成分分析では、変数を減らすことは難しいのではないかと考える

2. 次元リダクション

NTT COMWARE

（1）主成分分析の因子によるリダクション

前頁の主成分分析で求めた　因子数＝2　と　因子数＝3　を使用し、クラスタリングを実施した

因子数＝2

2

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	10	-2.7968	23.59
3	3	-2.836	31.67
4	10	-2.7859	39.57

因子数＝3

3

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	11	-4.1538	42.31
3	11	-4.131	58.26

予想通り、第１因子でデータのほとんどを説明できることから、それが影響している

　発現した　／　発現しない

という２つのクラスタに分かれたと思われる　現実的な解ではない結果　になった

２．次元リダクション

NTT COMWARE

（２）非線形回帰モデルによるリダクション

時間軸、ｄｏｓｅ軸、それぞれ

「遺伝子の反応はピークを迎えて減衰する」

という前提のもとに、それぞれの軸を２次関数で回帰させるモデルを考えることにより、
（４ｘ４＋１）＝１７次元を

ｆ（ｔ）　ｘ　ｇ（ｄ）　＋　Ｃ

```
f(t) = bt2 * t^2 + bt1 * t
g(d) = bd2 * d^2 + bd1 * d + bd0
```

という（２ｘ２＋１）＝５次元にリダクションするアプローチ。上記５つの係数を求める

以下の問題については、次のように対処した。

（１）観測誤差の取り扱い
（２）回帰式に当てはめたときに、回帰式適用後の発現量が負になるときの扱い
（３）非線形モデルを算出する統計ソフトウェア

（１）無視した
（２）無視した
（３）Ｓ言語を扱える　Ｓ－Ｐｌｕｓ２０００　でリダクションした

2．次元リダクション

NTT COMWARE

（2）非線形回帰モデルによるリダクション

前頁の非線形回帰モデルの係数に対して、主成分分析を実施したあとで、クラスタリングを実施した

主成分分析結果

因子	固有値	累積寄与率
因子 1	2.11	42.12%
因子 2	1.26	67.27%
(因子 3)	0.86	84.43%
(因子 4)	0.49	94.21%
(因子 5)	0.29	100.00%

クラスタリング実施結果

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	17	-2.4511	22.90
3	19	-2.3135	30.63
4	18	-2.2822	38.56

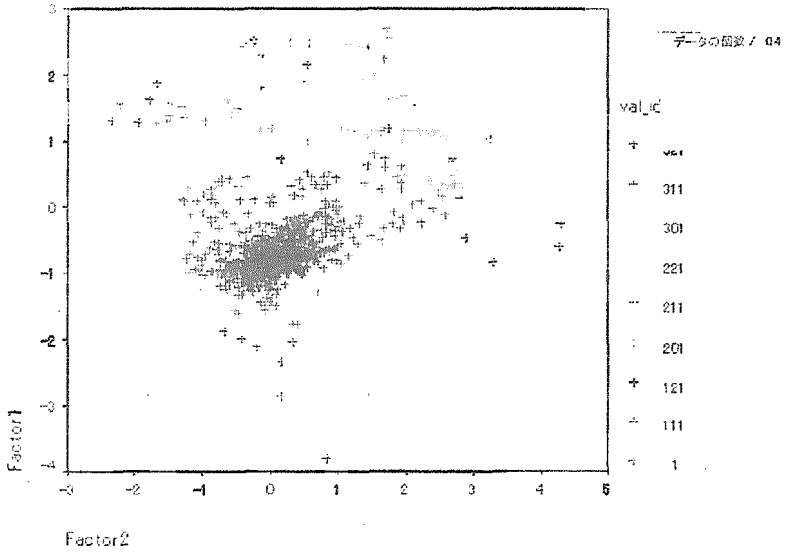
２つのクラスタに分かれた　現実的な解ではない結果　になった

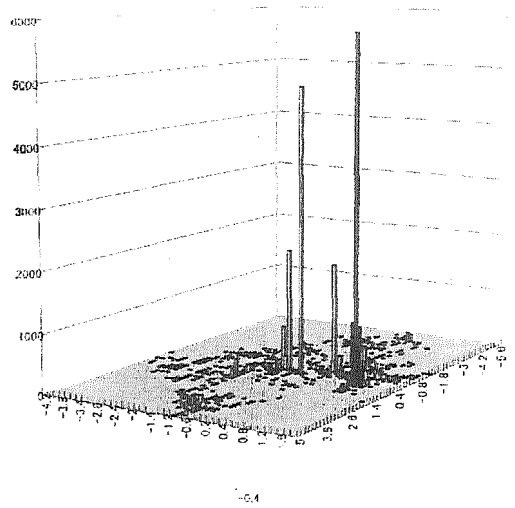
2．次元リダクション

（2）非線形回帰モデルによるリダクション

ＴＴＧ８について、全遺伝子の関数リダクションを実施し、ＥＭ法を実施した。

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	16	-2.4221	22.84
3	38	-0.5355	27.07
4	12	-2.3638	38.73





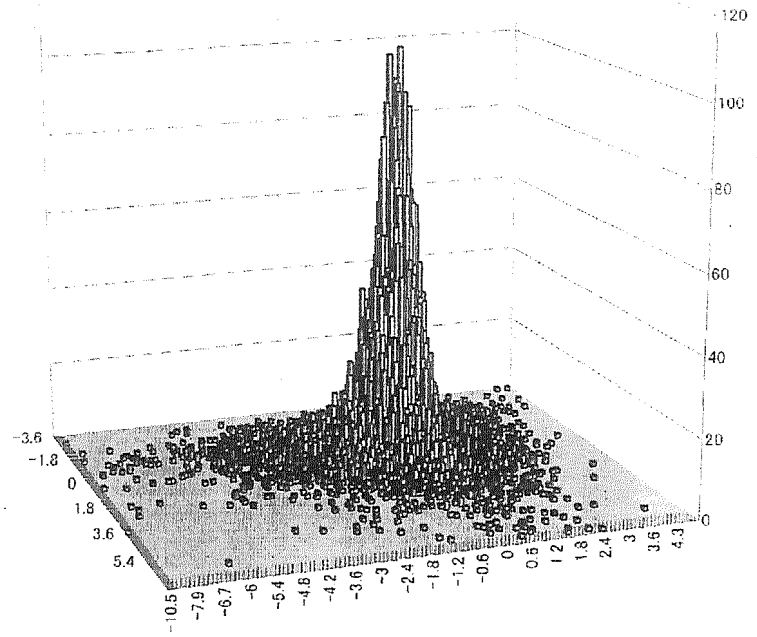
２つのクラスタに分かれた　現実的な解ではない結果　になった

2．次元リダクション

（3）矩形波を使った関数リダクションによるアプローチ検討

ＴＴＧ９について、全遺伝子を矩形波で変換した後に、主成分分析を行い、ＥＭクラスタリングを実施した。

クラスタ数	収束ステップ数	最大対数尤度	AIC
2	8	−7.53	121.06
3	15	−7.43	172.86
4	20	−7.37	224.74



２つのクラスタに分かれた　現実的な解ではない結果　になった