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

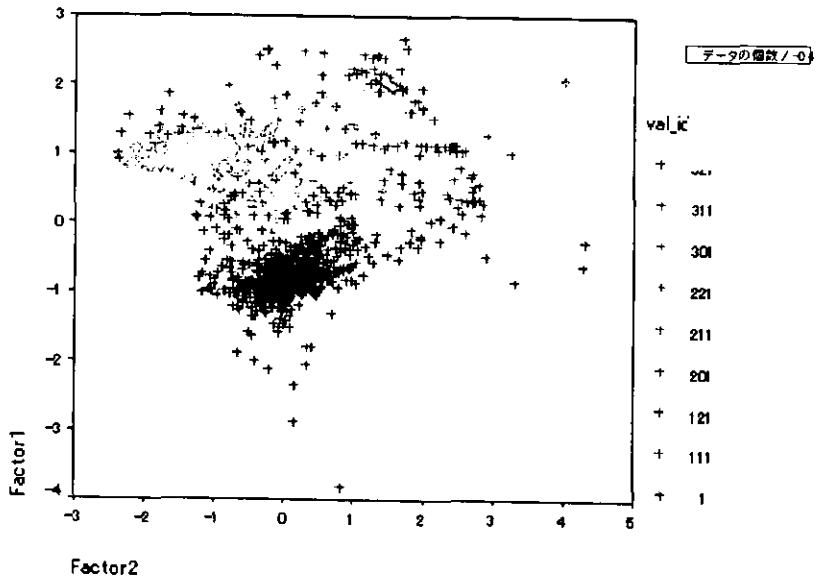
2．次元リダクション

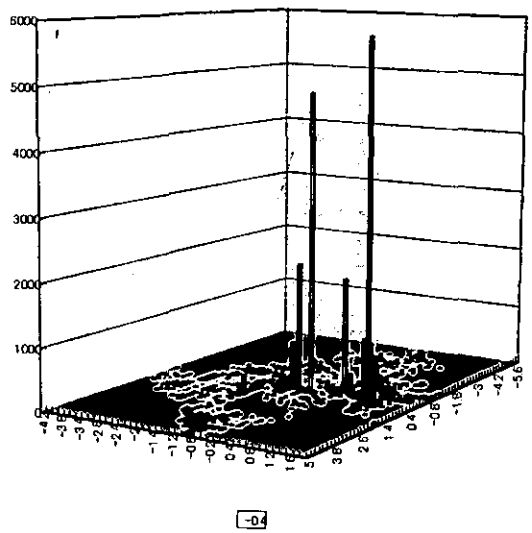
NTT COMWARE

（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

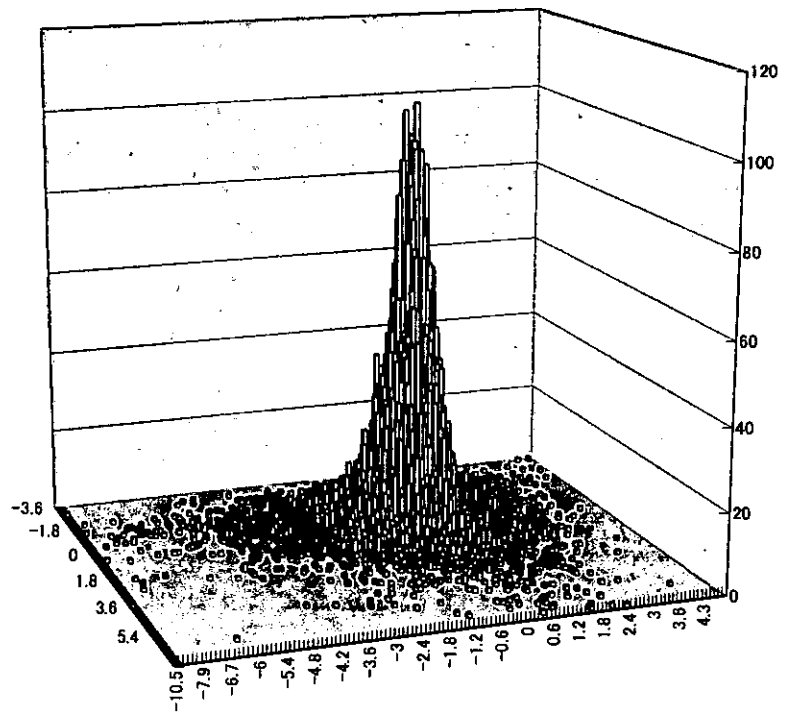
2．次元リダクション

NTT COMWARE

（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



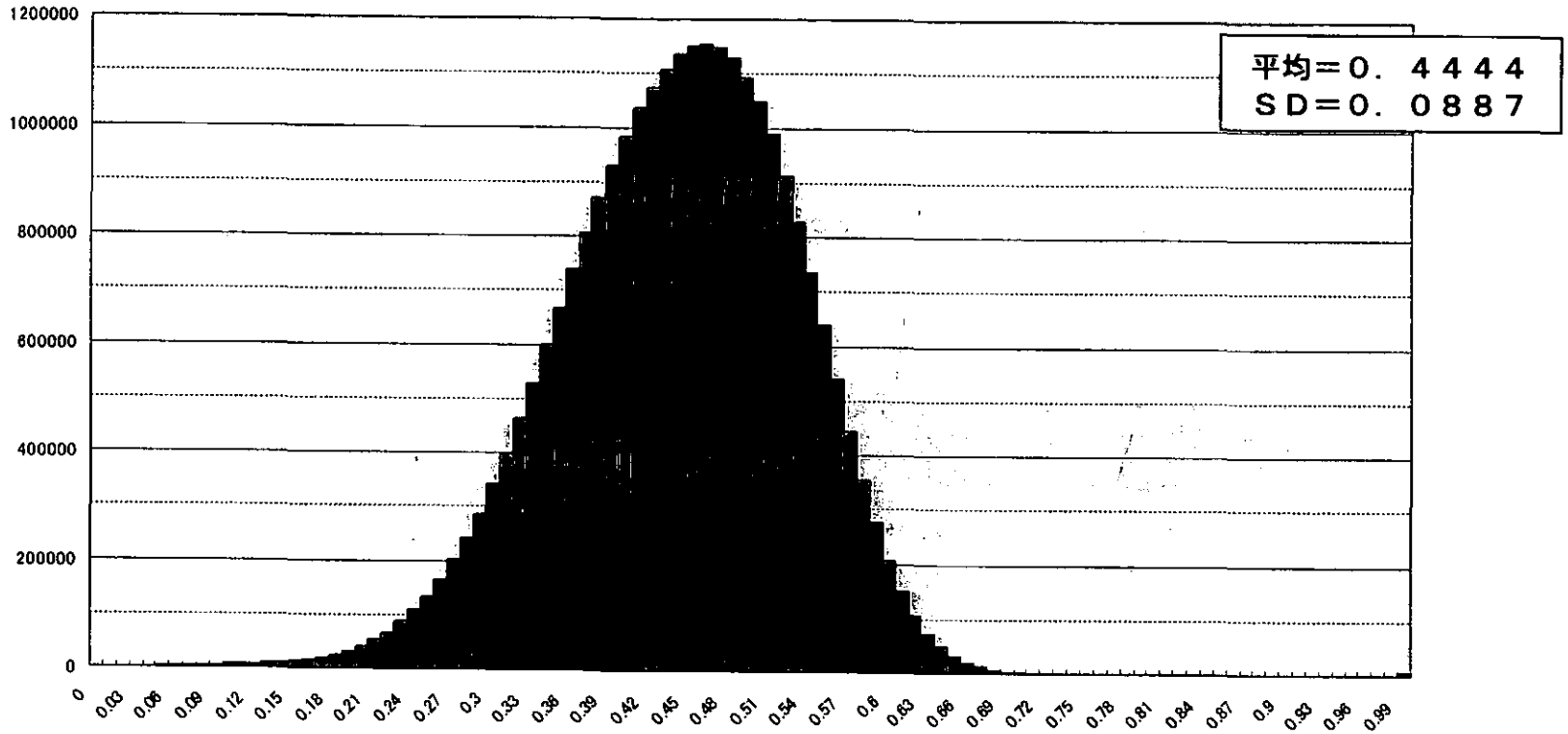
２つのクラスタに分かれた　現実的な解ではない結果　になった

３．ＴＭＦ計算

NTT COMWARE

（１）ＴＭＦ計算結果の分布

貴研究所、ご提示の遺伝子間の類似度計算方法（TMF）を全組み合わせについて、計算しその分布を示す。



正規分布のような釣鐘型の形状となっている。TMFの特性上、このような形状が保障されるわけではないが、概ねこのような形状になると類推される。

３．ＴＭＦ計算結果報告

NTT COMWARE

（２）性能考察

様々な改良を施した結果、以下のような時間となった

（１）ＴＴＧ９，ＴＴＧ１０　５，１３２Ｇｅｎｅ　：　３時間３０分

レコード件数＝５１３２ｘ５１３２÷２　・・・　約１３１７万

（２）ＵＴＥＲＯ　６，９７６Ｇｅｎｅ　：　７時間３０分

レコード件数＝６９７６ｘ６９７６÷２　・・・　約２４３３万

（３）ＵＴＥＲＯ全件　１２，４８８Ｇｅｎｅ　：　１５時間３０分

レコード件数＝１２４８８ｘ１２４８８÷２　・・・　約７７９８万

（３）ＴＴＧ９全件　２２，６９０Ｇｅｎｅ　：　５８時間３０分

レコード件数＝２２６９０ｘ２２６９０÷２　・・・　約２億５７４２万

ただし、Ｔｅｒａｄａｔａは別ＰＪと共同で使用しているため、クリアな状態ではない

４．密度ベースと階層的クラスタリングの融合

NTT COMWARE

（１）Ｚスコアを用いた階層的クラスタリング

Ｚスコアを用いた、全Ｇｅｎｅによる階層的クラスタリングについて取り組んだ

（１）前提条件

全Ｇｅｎｅ間の類似度を表すために、何らかの類似度（または非類似度）を定義できたと仮定する

（２）前処理

・全Ｇｅｎｅについて、全ｄｏｓｅ／時間を対象にＺスコアを計算する
・上で計算した１６個のＺスコアを、１６次元空間の座標とみなす
・全Ｇｅｎｅについて、ユークリッド距離で類似度を計算する（※）

※２つのＧｅｎｅ、Ａ，Ｂを与えたときに、対称性があることから　Ａ≧Ｂという制約をつける

・クラスタリング手法

閾値ごとに、DBSCANクラスタリング手法（改良版）を実施し、その結果を用いて、階層的な表現とする

・問題点

・閾値を決定するのが難しい
・ノイズの影響により、クラスタリングの結果に大きく影響を与える可能性がある

４．密度ベースと階層的クラスタリングの融合

NTT COMWARE

（２）ＴＭＦ計算結果を用いた階層的クラスタリング

TMF計算結果を用いた、階層的クラスタリングを改良し、５１２９Ｇｅｎｅで実施した

（１）前提条件

Ｚスコアと同じ

（２）処理概要

Ｚスコアと同じ

（３）クラスタリング結果（閾値＝０．７２　ｎ＝４）

１４４コアＧｅｎｅ　：　４６クラスタ

※問題点

Ｚスコアと同様に、閾値とｎの値を決定するのが難しい

4．密度ベースと階層的クラスタリングの融合

NTT COMWARE

（2）ＴＭＦ計算結果を用いた階層的クラスタリング（クラスタ結果）

TMF計算結果を用いた、階層的クラスタリングを改良し、５１２９Ｇｅｎｅで実施した

id	clst_label	gene_core		
1	1415682_at	1415682_at		
2	1415699_a_at	1415699_a_at	1420524_a_at	1433552_a_at
		1415702_a_at	1421847_at	1434646_s_at
		1415813_at	1421853_at	1436046_x_at
		1415819_a_at	1423077_at	1437503_a_at
		1415948_at	1423416_at	1437859_x_at
		1415961_at	1423912_at	1438192_s_at
		1416252_at	1423939_a_at	1448365_at
		1416635_at	1424311_at	1448495_at
		1416742_at	1424461_at	1448644_at
		1416764_at	1424488_a_at	1448846_a_at
		1417157_at	1425129_a_at	1450746_at
		1417390_at	1426111_x_at	1451357_at
		1417544_a_at	1426233_at	1451362_at
		1417744_a_at	1426783_at	1451782_a_at
		1417785_at	1427078_at	1452052_s_at
		1418112_at	1427720_a_at	1452147_at
		1418522_at	1428843_at	1455001_x_at
		1418794_at	1429296_at	1455855_x_at
		1419260_a_at	1429534_a_at	1456059_at
		1419445_s_at	1429707_at	1456313_x_at
		1419819_s_at	1433461_at	

id	clst_label	gene_core
3	1415867_at	1415867_at
		1416284_at
		1449618_s_at
		1450275_x_at
		1450735_at
		1452164_at
		1452354_at
4	1415870_at	1415870_at
		1416153_at
		1417538_at
		1419950_s_at
		1420846_at
		1423615_at
		1426824_at
		1448192_s_at
		1452051_at
5	1415876_a_at	1415876_a_at
6	1415979_x_at	1415979_x_at
		1416219_at
		1416937_at
		1438477_a_at
		1451068_s_at

id	clst_label	gene_core
7	1415986_at	1415986_at
8	1416215_at	1416215_at
9	1416372_at	1416372_at
10	1416500_at	1416500_at
		1416867_at
		1417285_a_at
		1424235_at
		1426162_a_at
		1435395_s_at
		1448697_s_at
11	1416595_at	1416595_at
12	1416671_a_at	1416671_a_at
13	1416789_at	1416789_at
		1417468_at
14	1417508_at	1417508_at
15	1417934_at	1417934_at
16	1418631_at	1418631_at
		1419177_at
17	1418897_at	1418897_at
18	1418998_a_at	1418998_a_at
		1428631_a_at
		1448823_at

id	clst_label	gene_core
19	1420460_a_at	1420460_a_at
20	1420525_a_at	1420525_a_at
21	1422487_at	1422487_at
		1452691_at
22	1422576_at	1422576_at
23	1423394_at	1423394_at
24	1423739_x_at	1423739_x_at
		1424041_s_at
		1449710_s_at
25	1424008_a_at	1424008_a_at
		1424406_at
26	1424039_at	1424039_at
		1428138_s_at
		1452055_at
27	1424117_at	1424117_at
28	1424694_at	1424694_at
29	1424708_at	1424708_at
30	1426414_a_at	1426414_a_at
31	1426679_at	1426679_at
32	1427060_at	1427060_at
33	1427896_at	1427896_at
		1451700_a_at

id	clst_label	gene_core
34	1428218_a_at	1428218_a_at
35	1431423_a_at	1431423_a_at
36	1431431_a_at	1431431_a_at
37	1434251_at	1434251_at
38	1435995_at	1435995_at
39	1438159_x_at	1438159_x_at
40	1448206_at	1448206_at
41	1448279_at	1448279_at
42	1448821_a_at	1448821_a_at
		1450431_a_at
43	1450720_at	1450720_at
		1455152_at
44	1452683_at	1452683_at
45	1452917_at	1452917_at
46	1454955_at	1454955_at

4．密度ベースと階層的クラスタリングの融合

NTT COMWARE

（2）ＴＭＦ計算結果を用いた階層的クラスタリング（クラスタ情報）

id	clst_label	min_gene	n	tmf_val
1	1415682_at	1415682_at	1	1.0000
2	1415699_a_at	1426233_at	62	0.5708
3	1415867_at	1416284_at	7	0.5917
4	1415870_at	1416153_at	9	0.6164
5	1415876_a_at	1415876_a_at	1	1.0001
6	1415979_x_at	1438477_a_at	5	0.7111
7	1415986_at	1415986_at	1	1.0000
8	1416215_at	1416215_at	1	1.0000
9	1416372_at	1416372_at	1	0.9999
10	1416500_at	1448697_s_at	7	0.6705
11	1416595_at	1416595_at	1	0.9999
12	1416671_a_at	1416671_a_at	1	0.9999
13	1416789_at	1417468_at	2	0.7280
14	1417508_at	1417508_at	1	1.0001
15	1417934_at	1417934_at	1	1.0000

id	clst_label	min_gene	n	tmf_val
16	1418631_at	1419177_at	2	0.7238
17	1418897_at	1418897_at	1	1.0001
18	1418996_a_at	1448823_at	3	0.6623
19	1420460_a_at	1420460_a_at	1	1.0000
20	1420525_a_at	1420525_a_at	1	1.0000
21	1422487_at	1452691_at	2	0.7474
22	1422576_at	1422576_at	1	1.0000
23	1423394_at	1423394_at	1	1.0001
24	1423739_x_at	1449710_s_at	3	0.7140
25	1424008_a_at	1424406_at	2	0.7278
26	1424039_at	1428138_s_at	3	0.6615
27	1424117_at	1424117_at	1	1.0001
28	1424694_at	1424694_at	1	1.0000
29	1424708_at	1424708_at	1	1.0000
30	1426414_a_at	1426414_a_at	1	1.0000

id	clst_label	min_gene	n	tmf_val
31	1426679_at	1426679_at	1	1.0001
32	1427060_at	1427060_at	1	0.9999
33	1427896_at	1451700_a_at	2	0.7241
34	1428218_a_at	1428218_a_at	1	1.0001
35	1431423_a_at	1431423_a_at	1	1.0000
36	1431431_a_at	1431431_a_at	1	1.0000
37	1434251_at	1434251_at	1	1.0000
38	1435995_at	1435995_at	1	1.0000
39	1438159_x_at	1438159_x_at	1	1.0001
40	1448206_at	1448206_at	1	1.0000
41	1448279_at	1448279_at	1	1.0001
42	1448621_a_at	1450431_a_at	2	0.7364
43	1450720_at	1455152_at	2	0.7400
44	1452683_at	1452683_at	1	0.9999
45	1452917_at	1452917_at	1	1.0000
46	1454955_at	1454955_at	1	1.0000

４．密度ベースと階層的クラスタリングの融合

NTT COMWARE

（３）ｄを動かしたときの階層的クラスタリングの違いについて

ＴＭＦクラスタリングに関して、ｄ＝３　と　ｄ＝４　について、簡単に比較してみました

下記にｄ＝３とｄ＝４におけるクラスタ数と出現Ｇｅｎｅ数についてまとめています

	dens	0.700	0.705	0.710	0.715	0.720	0.725	0.730	0.735	0.740	0.745
N of Clst	3	220	212	197	171	154	128	104	83	62	45
N of Gene		3575	2723	2015	1453	1021	667	438	286	179	120
N of Clst	4	91	84	55	69	51	52	44	35	28	14
N of Gene		2348	1723	1186	814	518	330	200	126	78	48

	dens	0.750	0.755	0.760	0.765	0.770	0.775	0.780	0.785	0.790	0.795
N of Clst	3	30	18	9	7	4	3	2	2	2	2
N of Gene		72	52	37	32	25	23	22	22	22	21
N of Clst	4	7	3	2	2	2	2	2	2	2	2
N of Gene		33	28	25	22	22	22	20	20	20	20

	dens	0.800	0.805	0.810	0.815	0.820	0.825	0.830	0.835	0.840	0.845
N of Clst	3	2	2	2	2	1	1	1	1	1	1
N of Gene		20	20	18	17	14	14	14	14	14	13
N of Clst	4	2	2	2	1	1	1	1	1	1	1
N of Gene		19	17	17	14	14	14	14	14	13	12

	dens	0.850	0.855	0.860	0.865	0.870
N of Clst	3	1	1	1	1	1
N of Gene		13	12	12	12	10
N of Clst	4	1	1	1	1	1
N of Gene		12	12	11	10	9

5．関数近似手法

NTT COMWARE

（１）前提条件

前提条件①

時間とドーズを次のように置き換える

時間	0h	2h	4h	8h	24h
T	0	1	2	3	4

投与量	0	0.1	1.0	10.0
S	0	1	2	3

前提条件②（近似関数の構成方法）

（１） ｔ＝０においてドーズに依存せず、一定

（２）今回のデータの最も自由度の高い次数で表現する

$$
\begin{aligned}
f(t,s) = \beta_{00} \quad & + \beta_{10}t \quad + \beta_{20}t^2 \quad + \beta_{30}t^3 \quad + \beta_{40}t^4 \\
& + \beta_{11}ts \quad + \beta_{21}t^2s \quad + \beta_{31}t^3s \quad + \beta_{41}t^4s \\
& + \beta_{12}ts^2 \quad + \beta_{22}t^2s^2 \quad + \beta_{32}t^3s^2 \quad + \beta_{42}t^4s^2 \\
& + \beta_{13}ts^3 \quad + \beta_{23}t^2s^2 \quad + \beta_{33}t^3s^3 \quad + \beta_{43}t^4s^3
\end{aligned}
$$

$\beta_{01}, \beta_{02}, \beta_{03}$は、条件(1)を満たすため、恒等的に0

５．関数近似手法

NTT COMWARE

（１）前提条件

ＡＩＣによる係数数推定

遺伝子によっては、ノイズの部分が強く、高次方程式による近似が適切でない場合もある。これらの方程式の選択をＡＩＣを用いて行う。

$$AIC = -2 \cdot \text{最大対数尤度} + 2 \cdot \text{パラメータ数}$$

$$\text{最大対数尤度} = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{n}{2}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_i)^2}$$

時間、投与量、各々、何次までと区切って、係数を求める。それぞれの方程式に対するＡＩＣを計算し、比較する。ＡＩＣは、小さい値のものが良好な結果であると認める。絶対値ではなく、一連の比較対象の相対値として比較を行う必要がある。

５．関数近似手法

NTT COMWARE

（２）関数近似手法を用いた分析

関数近似手法を行った後、相関関数を用いた時間軸の分析を実施しました

（１）関数近似実施

（２）クラスタリング実施
　　関数近似結果を表現値に戻し、ＺＳＣＯＲＥ化を実施
　　単純な３パターンを排除（フラット、時間で１次増加、時間で１次減少）

（３）クラスタ毎に中心変動を求める

（４）クラスタ同士の相互相関係数関数を求める

関数近似を行うことにより、誤差を排除することはできた。しかし、その結果を用いてクラスタリングを行うと、再現性の低い遺伝子ほど先にクラスタとしてまとまるという結果となった。

適切ではない結果　になった。このアプローチは、クラスタリングの前処理として用いるべきではない　と考えられる。

５．関数近似手法

NTT COMWARE

（３）関数近似手法の改善

単純な多項式による関数近似だけではなく、生物学的な制約条件を組み込んだ多項式を作成し近似を行う。近似結果として改善されていると考える。

① 投与量＝０の場合には、時間に依存せず発現量は一定である

② 投与量＝０の場合には、０ｈと２４ｈは同じ発現量になる

③ 全投与量において、時間が２ｈ以下では、同じ発現量になる

④ 全投与量において、時間が４ｈ以下では、同じ発現量になる

⑤ 全投与量において、時間が８ｈ以下では、同じ発現量になる

生物学的制約は、手法自身は有効な手段と考えられる。しかし、新規の制約を組み込む際に数式への展開が必要なため、研究者による自由な拡張が制限されるため、分析手法として活用しにくいものと考えられる。