公募研究：2000～2004年度

# 知識情報処理的手法を用いたDNAチップからの遺伝子ネットワーク解析

●花井　泰三

九州大学農学研究院

**〈研究の目的と進め方〉**

最近、DNAチップが開発されており、遺伝子の機能を明らかにするための最も重要な手段として考えられている。外部環境の変化などに対して、遺伝子の発現情報を経時的に分析し、これらを解析することで、遺伝子の相互関係を明らかにし、遺伝子全体のネットワークが明らかになれば、未知の遺伝子の機能予測や、遺伝子発現のコントロール、薬剤の開発などに大きなインパクトがあるであろう。しかし、チップから得られる遺伝子発現情報は非常に多くの遺伝子の情報が一度に得られ、これら遺伝子は相互に抑制や活性化しあっており、非常に複雑な関係があると予想される。このため、従来の統計的手法では解析が困難であると考えられる。一方情報処理分野では、知識情報処理と呼ばれる複雑な関係をモデル化する方法が種々開発されており、申請者はその有効性をこれまで明らかにしてきた。この方法を用いて多くの遺伝子発現情報を解析する。

本研究では、開発する手法の正当性を確認するために、すでに遺伝子の発現調節機能がほぼわかっている、酵母の分裂過程に的を絞り、DNAチップによる遺伝子発現データの取得を行う。DNAチップにより実験誤差の少ない定量PCRによる実験も、主な働きをする２０から３０の遺伝子に対して行う。この実験とコンピュータ解析を行い、どの遺伝子が転写制御を受けているのかを調べ、遺伝子ネットワークを明らかにし、既に知られている生化学的事実と照らし合わせて、本手法の正当性を確認する。このように計算手法の正当性確認のための実験を行うことで、より実験科学者を説得できると思われる。

そこで本研究では、酵母分裂時の、遺伝子発現データをDNAチップまたは定量PCR法を用いて経時的に計測し、知識情報処理の手法を用いて解析を行い、遺伝子ネットワークを明らかにすることを目指している。そのため、下図に示す三項目に関する研究を行う。

①適応共鳴理論（ART）など新規クラスタリング手法による遺伝子発現情報のクラスタリング

②人工ニューラルネットワーク(ANN)、ファジィニューラルネットワーク(FNN)による遺伝子発現量予測モデルの構築

③構築したANN、FNNの解析による遺伝子ネットワークの解析

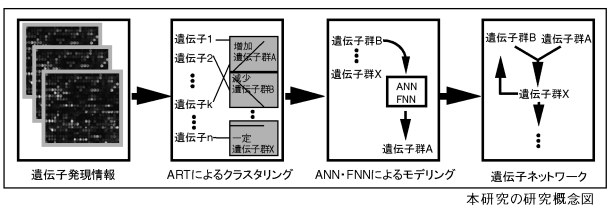



Fig.1 本研究の模式図

**〈研究開始時の研究計画〉**

①DNAチップによる酵母分裂時の遺伝子発現データの採取

分裂酵母を同調培養し、酵母分裂時の遺伝子発現データを同研究院保有のDNAチップ解析装置を用いて経時的に採取する。同時に、酵母分裂時に重要な働きを持つ遺伝子に関して、定量PCR装置による測定も行う。この際サンプリング間隔をできるだけ細かくし、どの程度のデータ点数であれば、この後の解析に十分であるかの検討も行う。

②適応共鳴理論(ART)などによるクラスタリング

経時的に採取した遺伝子発現データの時間的なパターンに従って、クラスタリングを行う。クラスタリングには従来から行われている、統計的手法のクラスタ分析と、ARTによる解析を行う。ARTは教師信号なし学習によってクラスタリングを行うアルゴリズムである。まずは、データベースのあった値を使用し、各種パラメータによる影響や他種法との比較を行う。さらに、実際取得したデータにより解析を行う。また、これらの結果得られたクラスタ群の情報と既に知られている遺伝子の情報と比較することで、本手法の有効性を確認する。

③人工ニューラルネットワーク(ANN)およびファジィニューラルネットワーク(FNN)による遺伝子発現モデルの構築

あるクラスタの経時的遺伝子発現パターンを他のクラスタの発現パターンから予測するANNおよびFNNモデルを構築する。あるクラスタの発現パターンを予測するために重要なクラスタを選別するにあたって、変数増加法または遺伝的アルゴリズムを使用する。この結果ある遺伝子群が発現する際に関与する遺伝子群が明らかになろう。さらにFNNを用いたモデルからは、どの遺伝子群がどのように遺伝子発現に関与しているかもIF-THENルールの形で得ることができる。また、これらの結果得られた情報と既に知られている知見と比較することで、本手法の有効性を確認する。

④構築したANNおよびFNNによる遺伝子ネットワークの解析

個々の遺伝子発現クラスタを予測するモデルをそれぞれネットワークとして接続し、分裂の際の、遺伝子発現の様子をシミュレートする。

⑤他の細胞や刺激による遺伝子発現の誘導とシミュレーション

酵母以外にも他の微生物や、動物細胞などを様々な条件かで遺伝子の誘導などを行い、これらの実験条件下での遺伝子発現のデータも追加採取する。それらに対しても遺伝子発現の状況をシミュレートし、遺伝子ネットワークの構築とシミュレーションを行う。

**〈研究期間の成果〉**

①および⑤　DNAチップおよびマイクロアレイの特性を調べるため、大腸菌、酵母、ヒト細胞に、刺激を与え、その際のmRNA変化を測定した。また、定量PCRを用いて、同様の測定を行い、誤差、再現性について検討を行った。

②酵母の胞子形成時の時系列マイクロアレイデータに対して、従来法の階層的クラスタリング、k平均アルゴリズム、自己組織化ネットワークと今回あらたに適用するART（適用共鳴理論）による解析を行った。その結果、生化学的知見に基づく結果に最も近い結果が得られたのは、ARTであった。いくつかの遺伝子が、従来の地検とは異なるクラスタリング結果になったので、これらの遺伝子を生化学的に詳細に調べたところ（Timeで書かれた発現時期が同じ遺伝子が同じクラスタに属した方がよいと考えられる）、今回解析に使用した実験条件下では、Fuzzy ARTの結果を支持するものとなった。得られたクラスタの各サンプリングポイントにおける、データの分散も、Fuzzy ARTによる結果が最も小さいものであった。また、マイクロアレイによる実験は多くのノイズを含むため、実際の実験データに人工的にノイズを加えたデータを作成し、前出の手法を適応したところ、ARTによって形成された遺伝子のグループは、ノイズを加えない場合とほぼ同じ結果が得られたが、他の手法を用いた場合、遺伝子のグループは大きく異なるものとなった。この結果は、5組のランダムにノイズを加えたデータを作成し、繰り返しクラスタリングを行った場合でも、同様の結果が得られた。さらに、得られたクラスタリング結果と異なる傾向の発現パターンをもつ遺伝子のクラスタリングを行ったところ、得られたクラスタ以外に新たなクラスタを形成し、本手法の有効性を示すことができた。

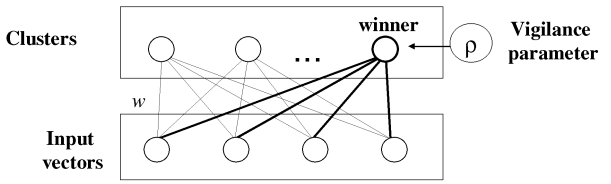



Fig.2 Fuzzy ARTの模式図

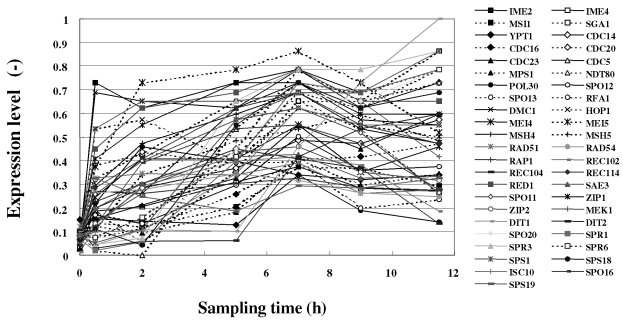



Fig.3 解析に用いた胞子形成DNAマイクロアレイデータ

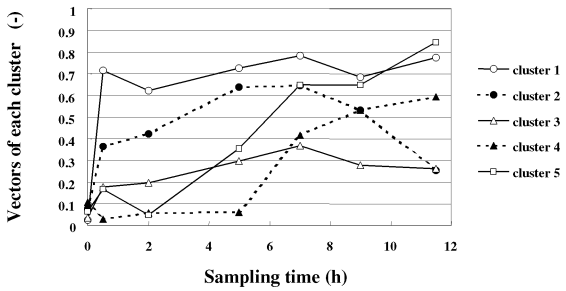



Fig.4 Fuzzy ARTによるクラスタリング結果

Table 1 Fuzzy ARTでのクラスタリング結果

| Cluster | Name | Time |
|---|---|---|
| 1 | *DMC1* | Early |
| | *IME2* | Early |
| | *MEI5* | |
| | *RED1* | Early |
| 2 | *HOP1* | Early |
| | *MEK1* | Early |
| | *MSH4* | |
| | *MSH5* | |
| | *REC114* | Early |
| | *SPO11* | Early |
| | *SPO13* | Early |
| | *SPO16* | Early |
| | *ZIP1* | Early |
| 3 | *CDC14* | |
| | *CDC23* | |
| | *IME4* | Early |
| | *MEI4* | Early |
| | *MPS1* | |
| | *MSI1* | |
| | *POL30* | |
| | *RAD51* | |
| | *RAD54* | |
| | *RAP1* | |
| | *REC102* | Early |
| | *REC104* | Early |
| | *RFA1* | |
| | *SAE3* | |
| | *SPO12* | Middle |
| | *SPS19* | |
| | *YPT1* | |
| | *ZIP2* | |
| 4 | *CDC16* | |
| | *DIT1* | Mid-Late |
| | *DIT2* | Mid-Late |
| 5 | *CDC20* | |
| | *CDC5* | |
| | *ISC10* | |
| | *NDT80* | |
| | *SGA1* | Late |
| | *SPO20* | |
| | *SPR1* | Late |
| | *SPR3* | Late |
| | *SPR6* | |
| | *SPS1* | Middle |
| | *SPS18* | |

Table 2 各クラスタリング手法による正答率（同じクラスタに異なるTimeの遺伝子が入っていると不正解とする）

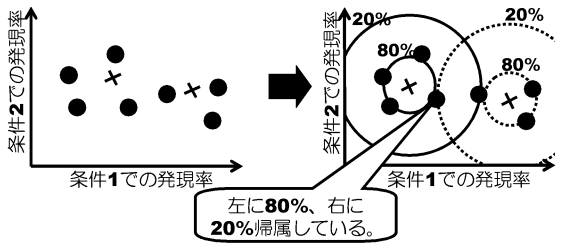
| | |
|---|---|
| Fuzzy ART | 0.90 |
| Hierarchical clustering | 0.81 |
| k-means clustering | 0.86 |
| SOMs | 0.86 |

Table 3 各クラスタリング手法による再現率（人工的なノイズを5パターン付加し、ノイズなしのクラスタリング結果と一致した平均%）

| | |
|---|---|
| Fuzzy ART | 79.10 |
| Hierarchical clustering | 73.30 |
| k-means clustering | 55.60 |
| SOMs | 57.30 |

Fuzzy ARTクラスタリング以外にも、k平均クラスタリングにFuzzy推論を導入したFuzzy k平均クラスタリングについても研究を行い、その特性がFuzzy推論を導入していない、k平均クラスタリングより優れていることを明らかにした。特に、人工的なノイズを大きさを変えて追加した際に、k平均クラスタリングでは、小さいノイズから再現率は低いものであったが、Fuzzy k平均クラスタリングでは、ノイズが小さい場合はノイズがない場合と同じクラスタリング結果が得られ、ノイズを大きくした場合でも、8割以上でノイズがない場合と同じ結果が得られることが分かった。ノイズが多いマイクロアレイデータの解析法としては、非常に優れたクラスタリング手法であることが明らかになった。

また、遺伝子の発現開始および分解開始に注目したMathematical model based clustering(MMBC)と呼ぶ手法を世界で初めて開発した。時系列データのクラスタリングでは、時間的に連続しているデータを扱うにもかかわらず、各サンプリングポイントは全く関連のない別の軸のデータとして取り扱っている。しかし、別々に測定されたデータと比較し、時間的に連続したデータには動的な変化を含んでいると考えられ、これらの情報を有効に活用することが望まれる。




遺伝子は各クラスタへの帰属度の和が
**100**%となるように複数のクラスタに帰属する。

Fig.5 Fuzzy k平均クラスタリングの模式図

Table 4 k平均クラスタリングとFuzzy k平均クラスタリングの再現率

| k-means クラスタリング | | |
|---|---|---|
| ノイズの大きさ(%) | 同じクラスタリング結果が得られた遺伝子数 / 全遺伝子数 | 再現率(-) |
| 50 | 29 / 45 | 0.644 |
| 100 | 29 / 45 | 0.644 |
| 200 | 28 / 45 | 0.622 |

| Fuzzy k-means クラスタリング | | |
|---|---|---|
| ノイズの大きさ(%) | 同じクラスタリング結果が得られた遺伝子数 / 全遺伝子数 | 再現率(-) |
| 50 | 45 / 45 | 1.000 |
| 100 | 45 / 45 | 1.000 |
| 200 | 38 / 45 | 0.844 |

マイクロアレイデータのクラスタリングを行う際、**Fuzzy k-means**クラスタリングは **k-means**クラスタリングよりノイズに強い。

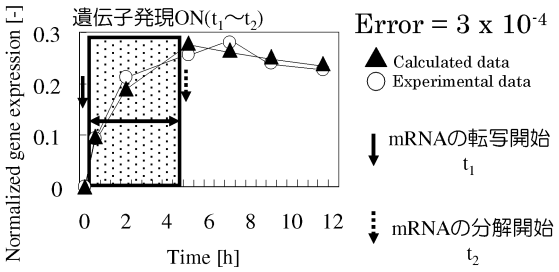



Fig.6 MMBC法に適応する数学モデルの実験データへの適応例

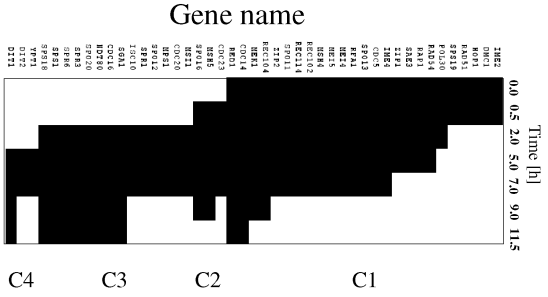



Fig.7 数学モデルで最適化された結果、遺伝子発現している時間

そこで、遺伝子発現を表す単純な数学モデルを構築し、測定データを再現するようにこの数学モデル内のパラメータを最適化した。これら最適化後のパラメータをクラスタリングすることで、時間的に連続している情報を有効活用することを試みた。適応を試みた数学モデルの2つのパラメータに注目し、それら二つの値を用いてクラスタリングを行った。クラスタリングには最も一般的なk平均クラスタリングを用いた。パラメータをクラスタリングした結果と、そのままサンプリング点をクラスタリングした結果を比較したところ、パラメータをクラスタリングしたMMBCは、ノイズに強く、データがその生物

学的特徴に従い、はっきりと分かれることが明らかになった。

③遺伝子ネットワーク解析手法について、研究開始時の予定では、人工ニューラルネットワークやファジィニューラルネットワークを階層状に結合するネットワーク解析方法の構築を行う予定であったが、最適化すべきパラメータが莫大で、これを最適化するよい方法をないことから、別の方法を選択することとした。研究開始時と比較し、この分野は広く研究されるようになり、現在では、大きく分けてBoolean algorithm、Bayesian algorithm、S-system を用いた方法がある。これらの方法の中でもBoolean algorithmは、二値の情報間の因果関係を推定する方法であり、アルゴリズムや概念が理解しやすいことから、今後この分野で広く利用されることが予想されたため、この方法を用いた解析を行うこととした。

Boolean algorithmでDNAチップデータの解析を行うためには、連続値から二値への変換を行う必要がある。二値化には閾値を用いることとなるが、現在までのところ、Boolean algorithmの閾値に関して十分な研究は行われておらず、解析者が恣意的に決定しているのが現状であり、何らかの最適化手法の開発が望まれている。そこで、我々はKyoto Encyclopedia of Genes and Genomes (KEGG)などの既存の遺伝子やタンパク質の相互作用に関する生物学情報を用いて、閾値の最適化を試みた。つまり、KEGGなどに記載されている情報とBoolean algorithmによって推定される相互作用の一致率が最も高くなるような、閾値を求めることとした。解析に用いたデータは酵母の細胞周期に関するmRNAの経時変化を用いた。

本手法の有効性を確認するために、様々なタイプの仮想的な遺伝子ネットワークシミュレーションモデルを作成し、ここから得られる時系列発現データの解析を行ったところ、ほぼすべての場合で高い推定精度を示した。さらに上記の細胞周期データの解析を行ったところ、KEGGに記述されるデータ以外の遺伝子またはタンパク質の相互作用が推定された。これらについて詳細に文献調査を行ったところ、そのいくつかの相互作用が証明された。以上のことから、本手法を用いることで、未知の相互作用推定の可能性が示唆された。

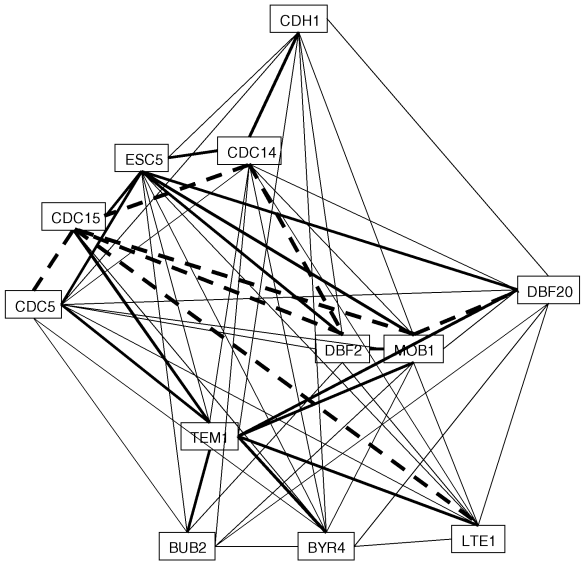



Fig.7 ブーリアンネットワークによって決定された遺伝子ネットワーク（点線がKEGGには存在せず、他の文献で存在が示唆された相互作用）

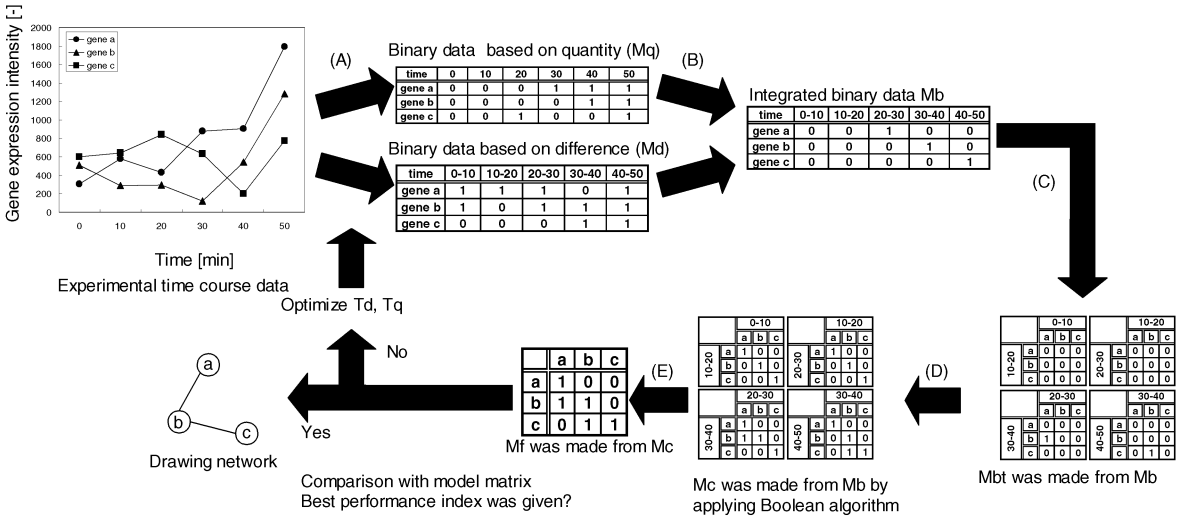



Fig. 6 既知遺伝子ネットワーク情報によるブーリアンネットワークの閾値最適化

**〈国内外での成果の位置づけ〉**

最近、世界中の様々なグループで、遺伝子ネットワーク解析用のソフトウェアが開発されている。しかし、実際の実験データを適応している例は少なく、我々のグループのように実験と情報処理を一緒に同じ場所で行う例はほとんどない。また、遺伝子のクラスタリングを行い、クラスタ同士の関係を用いて解析する例も、少ない。

クラスタリング解析に関しては、一つの研究室でこれだけ様々な方法で解析した例はなく、特にMMBC、Fuzzy ARTに関しては、世界的にもユニークな研究である。また、Boolean algorithmを用いた遺伝子ネットワークに関しても、実際の実験データを適応した際に問題になる閾値の問題に注目して、解析したことは大変意義深い。

**〈達成できなかったこと、予想外の困難、その理由〉**

クラスタリング解析に関しては、クラスタ数の決定法についての検討が、研究期間が不足しできなかった。また、遺伝子ネットワーク解析に関しては、我々が利用したBoolean algorithmとBayesian algorithm、S-system、GGMなどとの比較が必要であったと考えられる。何より、解析を対象とした例が少なかったため、他の生物を利用した場合、他の刺激を与えた場合どのような結果が得られるかなどの検討が必要であったと考えられる。

予想外の困難としては、予想以上に実験データの再現性が低く、計算する際に許容できるレベルにまで到達することが難しかった。実験データの誤差は通常25から50%程度を許容しているのに対して、アルゴリズム的には大きくても5から10%の誤差しか許容できない。将来の実験再現性向上を待つか、コストを度外視し多数こなすことで、統計的に誤差を少なくするしか、現状では、方法はないと思われる。

**〈今後の課題〉**

「達成できなかったこと」の部分でも書いたが、クラスタ数決定法、様々な遺伝子ネットワーク解析法の比較、および他のデータの適応が今後の課題であろう。

**〈研究期間の全成果公表リスト〉**

**1)論文**

1.
Taizo Hanai, Hiroyuki Hamada, Masahiro Okamoto: Application of bioinformatics for DNA microarray data to Bioscience, Bioengineering and medical field, Journal of Bioscience and Bioengineering, in press (2006).

2.
Kazumi Hakamada, Masahiro Okamoto, Taizo Hanai: Novel technique for preprocessing high dimensional time-course data from DNA microarray: mathematical model-based clustering, Bioinformatics, in press (2006).

3.
Chinatsu Arima, Kazumi Hakamada, Masahiro Okamoto, Taizo Hanai: Validity Index for Fuzzy K-means Clustering using The Gap Statistic Method, Genome Informatics, 16, P040 1-2 (2005)

4.
Kazumi Hakamada, Masahiro Okamoto, Taizo Hanai: Validation of Mathematical Model Based Clustering by using Time-course data, Genome Informatics, 16, P041 1-2 (2005)3.

5.
Kazumi Hakamada, Taizo Hanai, Masahiro Okamoto: Mathematical Model Based Clustering of Gene Expression, Genome Informatics, 15, P037 1-2 (2004)

6. 040406134
Yoshihiko Tashima, Taizo Hanai, Hiroyuki Hamada, Masahiro Okamoto, Kinetics Behavior of G1-to-S Cell Cycle Phase Transition Model, Genome Informatics, 14, 608-609 (2003).

7. 0404061352
Kazumi Hakamada, Taizo Hanai, Masahiro Okamoto, Clustering Method Based on Onset and Cessation of Gene Expression, Genome Informatics, 14, 330-331 (2003).

8. 0404061358
Chihoko Tago, Taizo Hanai, Masahiro Okamoto, Prognosis prediction by microarray gene expression using Support Vector Machine, Genome Informatics, 14, 324-325 (2003).

9. 0404061401
Chinatsu Arima, Taizo Hanai, Masahiro Okamoto, Gene Expression Analysis Using K-means Clustering, Genome Informatics, 14, 334-335 (2003).

10. 211111424
Tomida, S., Hanai, T., Honda, H. and Kobayashi, T.: Analysis of expression profile using fuzzy adaptive resonance theory, Bioinformatics, 18, 8, 1073-1083 (2002).

11. 2111111453
Hakamada, K., Hanai, T., Honda, H. and Kobayashi, T.: Identifying genetic network using experimental time series data by Boolean algorithm, 12, 272-273 (2001).

12. 2111111504
Ando, T., Hanai, T., Honda, H. and Kobayashi, T.: Prognostic prediction of lymphoma by gene expression profiling using FNN, 12, 247-248 (2001).

13. 2111111526
Tomida, S., Hanai, T., Honda, H. and Kobayashi, T.: Gene expression analysis using Fuzzy ART, 12, 245-246 (2001).

**2)特許**

14.
適応共鳴理論による発現遺伝子のクラスタリング　出願中