

要旨

特定遺伝子組み合わせは、癌の治療予後の死亡率を左右する。本研究はこの特定遺伝子組み合わせを探索する基礎研究として、自己組織化マップによる関連遺伝子分類法の提案及びその結果を解析するツールの開発を行い、提案方法による数値計算に基づいた遺伝子解析結果を報告する。

1. はじめに

日本人の主要な死因が感染症から生活習慣関連病へと移行する中で、特に癌は、1981年以來、日本人の死亡原因の第1位となり、疾病対策上の最重要課題として対策が進められてきた。その中で、遺伝子の組み合わせが、癌による死亡率、すなわち予後を左右することが判明している。この予後の推定は、膨大な数の遺伝子から予後に関連する遺伝子の特定に基づく。本研究では、公表されている遺伝子のデータから、発症した癌の死亡率が高いか、低いかをできる限り正確に判別できる分類モデルを自己組織化マップ(Self-Organizing Maps: SOM)により作成する方法を提案し、更にその解析ツールの開発によって予後に関連する遺伝子を特定することを目的としている。

2. 従来の研究

予後診療には疾患予後と治療予後の2種類がある。疾患予後は治療せずに放置した場合の生存可能期間を意味し、治療予後は特定の非ホジキンリンパ腫を特定の診療法を初回治療として行った場合、どの程度の生存期間が期待できるかを指す。癌患者の予後を知ることができれば、疾患予後は治療をしないので別問題であるが、治療予後ではその人にあった治療法を考慮しながら癌と向かい合うことを可能とできる。従って、癌になった患者の治療予後つまり死亡率の高低を分類することは非常に重要な問題となる。現在、このような分類を行う方法としては国際予後指標(International Prognostic Index: IPI)を使う方法が実際の臨床の場で盛んに使用されている。しかし、近年、DNAあるいはRNAの状態を定量的もしくは定性的に解析することのできるマイクロアレイ法により、遺伝子を大量に解析することができるようになった。また、その解析データと情報科学の方法を用いて予後を診断する方法も出現している。このような方法としてはニューラルネットワークによる方法であるk-ミーン法、バックプロバケーションニューラルネットワーク、階層クラスタリングなどが使用されている¹⁾。

3. SOM

SOM²⁾は1981年頃にT.Kohonenにより提唱された、入力パターン群をその類似度に応じて自律的に獲得し分類する、教師なし学習ニューラルネットワークの一つである。

本研究で採用するSOMは、図1に示すように、入力層に入力データが n 次元実数ベクトルである参照ベクトル集合 $\mathbf{X}=[\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N]$ として与えられ、競合層には、ニューロン集合 $\mathbf{U}=[U_{ij}]$ が $M \times M$ の格子状に配置される。入力ベクトル \mathbf{X} は全てのニューロンに提示され、2次元格子配列上の点 (i, j) にあるニューロン U_{ij} はその入力ベクトル \mathbf{X} に結合した結合係数ベクトル \mathbf{W}_{ij} を持っている。

3.1. SOMの分類学習アルゴリズム

以下にSOMのアルゴリズムを示す。

- a) 入力ベクトル数を N 、学習終了回数を T とし、 \mathbf{W}_{ij} を乱数により設定する。

- b) 入力ベクトルから1要素を選択し、そのベクトルを $\mathbf{x} \in \mathbf{X}$ とする。
c) 入力ベクトル \mathbf{x} と各結合係数 \mathbf{W}_{ij} とのユークリッド距離 L を求める。 L は式(1)より求める。

$$L = |\mathbf{x} - \mathbf{W}_{ij}| = \sqrt{\sum_{k=1}^n (x_k - W_{jk})^2} \quad (1)$$

ただし、 $\mathbf{x}^T = [x_1, x_2, \dots, x_n]$ および $\mathbf{W}_{ij} = [W_{ij1}, W_{ij2}, \dots, W_{ijn}]$ である。

- d) (1)式より、 L が最小となるニューロン U_{ij} を見つける。
e) U_{ij} の結合係数 \mathbf{W}_{ij} を式(2)に示す更新式により変更する。

$$\mathbf{W}_{ij}(t+1) = \mathbf{W}_{ij}(t) + \alpha(\mathbf{x}_j(t) - \mathbf{W}_{ij}(t)) \quad (2)$$

ただし、 t は学習回数とし、 α は式(3)により学習回数によって減少する定数とする。

$$\alpha = \alpha_0 \left(1 - \frac{t}{T}\right) \quad (3)$$

その後、 U_{ij} の近傍のニューロンも式(2)で更新する。近傍係数 d は式(4)により変更する。

$$d = d_0 \left(1 - \frac{t}{T}\right) \quad (4)$$

- f) $t=T$ となるまで繰り返し、各入力ベクトルと最も類似度の高いニューロンにその入力ベクトルのラベルを付ける。

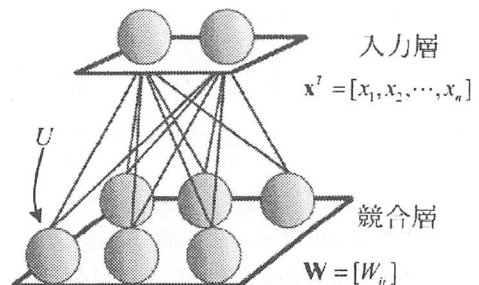


図1 SOMの原理図

4. SOMによる分類法

4.1. SOMによる分類

ウェブ上にあるびまん性大細胞型B細胞リンパ腫(DLBCL)の患者58人の7129種類の遺伝子データを使用し、SOMを用いて分類することで生存判別をする。このデータはマイクロアレイ法によって測定されたデータであり、各患者が5年後に生存しているか否かは、あらかじめ分かっている。従って5年以降で、生存している状態をクラス0(class0)、生存できない状態をクラス1(class1)とする。ついで、この遺伝子から分類に役立つものを前処理により選定し、SOMによる分類を実施する。

4.2. 遺伝子データの選定前処理

遺伝子データは、7129種類あるため、これらを全て利用すると計算量が膨大になる。また、全データがモデル構築に役立つとは限らないので必要なものを選定する必要がある。そこで、同じクラスでのデータのばらつきが少なく、違うクラスのデータとは格差がある、これら2点を満たす遺伝子を以下のように選定する。

各遺伝子に対して以下の2つの条件を満たすものを選定する。

$$|Max-Min| < 100 \wedge |Max/Min| < 3$$

ただし、ここでMax及びMinはWebに公開されている遺伝子の測定量である。

次に、上の条件を満たした遺伝子に対して、式(5)の評価 τ を適用する。

$$\tau = \frac{|\mu_0 - \mu_1|}{\sigma_0 + \sigma_1} \quad (5)$$

ここで、 μ は各クラスの平均値、 σ は各クラスの標準偏差である。式(5)では、クラス間の平均値の差が大きいほど、また同じクラスの中でばらつきが少ないほど絶対値が高くなる。以上より τ の高い遺伝子をもつSOMの入力(参照)ベクトルに利用する。

4.3. データの正規化

実際にモデルを構築するときには、データをそのまま利用するのではなく、何らかの前処理が必要である。これは、データの誤差の影響を少なくし、かつ正規化を行うためである。今回のSOMでは0から1までに値を正規化するため、各遺伝子に式(6)を適用する。

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (6)$$

この操作により、全遺伝子データが0から1までの数値として、正規化される。

5. 数値計算実験

以下に、数値計算実験の結果を示す。

5.1. SOMの適用

前処理を終えた各遺伝子を入力ベクトルの成分とし、58人分のデータを入力ベクトルとしてSOMを適用した。その分類結果を図2に示す。実験条件は表1の様に設定した。

表1 実験条件

入力ベクトル数	58個
入力ベクトルの次元数	100次元(遺伝子数)
ニューロン数	50×50個
学習回数	30000回
学習係数 α_0	0.25
近傍係数 d_0	30

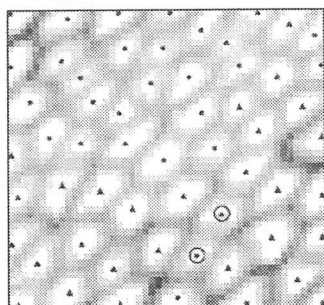


図2 SOMによる自己組織化分類

図2では、三角形のラベルがclass0の入力ベクトル、丸いラベルがclass1の入力ベクトルを示す。ここでは、隣り合っているラベルほど、その入力ベクトルの各要素の類似度が高い。更に、隣り合うものでもその境界線が薄いほど類似度が高くなる。これより、全体的には上下で2つのクラスに分類されることがわかる。しかし、中央下では三角形のラベルの中に丸いラベルがあるように、遺伝子のデータは似ているが、わずかな遺伝子の異なりによって生死が分かれている部分もあると予測される。またそれは、クラスの境界となっている中央部分でも同様のことが言える。

5.2. 差分解析

図2で、中央下にある丸で囲まれた入力ベクトルの今回SOMに使用した100種類の遺伝子のデータを図3~4に示す。

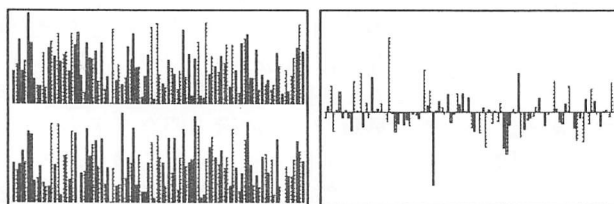


図3 遺伝子のデータ

図4 2つのデータの差

図3は、上がclass0のベクトル、下がclass1のベクトルの各データである。図4は、図3のデータから上のベクトルと下のベクトルの差をまとめたグラフである。図3~4からも、隣り合っているベクトルは基本的に各データが似ていることがわかる。しかし、明らかに違う部分も何点もあり、そこに生死を分ける要因があると予測することができる。

5.3. デンドログラムによる解析

図2の結果をもとにクラスター分析をして、デンドログラムを作成した。これを図5に示す。

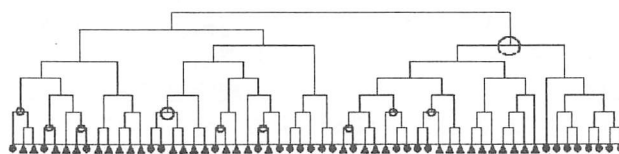


図5 デンドログラム

デンドログラムは、図2の結果から、ラベル間の距離が短いものを1つのものとして分類するものである。階層が下にある組み合わせほどベクトル同士の類似度が高いことを示す。図5から、丸で囲んだ部分の遺伝子間に予後に関する関連遺伝子があることが推定される。

6. おわりに

本研究で提案した方法と開発した解析ツールは、SOMに癌の予後診療に使える見込みがあることがわかった。より良い分類をするためには、GAなどの組み合わせが必要と考えられる。今後、具体的に癌の予後に関する遺伝子を特定するための解析ツールを更に開発する必要がある。また、別の手法による分析を行うか、公開されている結果を用いて本研究の結果を検証する必要がある。

参考文献

- 1) Margaret A. Sipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning", Nature Medicine, Volume 8, Number 1, January 2002
- 2) T. Kohonen 他; 自己組織化マップ, シュプリンガー・フェアラーク東京(1996)