

SOMによる癌の予後決定遺伝子の決定に関する基礎研究

旭川高専 ○木村 幸代, 渡辺 美知子, 古川 正志

要旨

これまでに, SOM を用いて5年後の癌予後遺伝子の特定に関するクラスターマップの作成を行い, その有効性を確認してきた. また, Ward 法を用いて遺伝子間の類似度表現を SOM で得られた患者の配置座標を実施してきた. 本研究では, SOM 自身が類似度に関する樹枝図表現を持つことを利用し, 樹枝図表現の自動生成法を提案し, 癌予後遺伝子の特定を行う基礎研究を報告する.

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 年以来, 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, 多数の遺伝子の中の特定の組み合わせと推定できる. これまでに, インターネット上で公表されている遺伝子のデータを採用し, 予後の発症を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てきた. 本研究では, この分類結果を基に, SOM から遺伝子の系統図を自動生成する方法を提案し, その結果とこれまでに Ward 法を用いて生成した遺伝子の分類樹枝図を検討する事で, 予後に関連する遺伝子の組み合わせを特定することを目的とする.

2. 従来の研究

近年, 実際の医療現場ではDLBCL等の国際予後指標 (International Prognostic Index, IPI) を利用した方法が盛んに用いられている. この予後推定法は, 医師が行う内診と同様で, 患者の年齢, 全身状態, 癌の状態などの項目から点数をつけ, その結果に応じて予後の分類を行うヒューリスティックな方法である. しかし, 近年では, マイクロアレイチップ法の開発により DNA や RNA に含まれる遺伝子を定量的もしくは定性的に測定し, こうした遺伝子情報を, 公開するようになった. このような遺伝子の大量の情報の解析方法として, ソフトコンピューティングを採用したバイオインフォマティクスによる予後推定法が研究されている. このような方法としては, ニューラルネットワークである K-ミーン法, サポートベクターマシン, 階層クラスタリングなどが採用されている¹⁾. 著者らは, 上記方法とはことなり, 遺伝子分類を視覚的に行える SOM による予後推定法を提案してきた. 本研究では, これまでの結果を基礎に, 予後の遺伝子を明らかにする解析ツールとして SOM に基づく遺伝子の系統図生成法を提案し, 今後 Ward 法による樹枝図の遺伝子分類と比較検討を行うことを目的とする.

3. SOM

3.1 SOM のモデル

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動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.

SOM は図 1 に示されるように入力層と競合層をもつニューラルネットワークの一つである. 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

$X=[X_1, X_2, \dots, X_N]$ が与えられ, 競合層には, ニューロン集合 $U=[U_{ij}]$ が予め設定したトポロジーの頂点群に配置される. 本研究ではこのトポロジーとして図 2 に示すヘキサゴンを採用した.

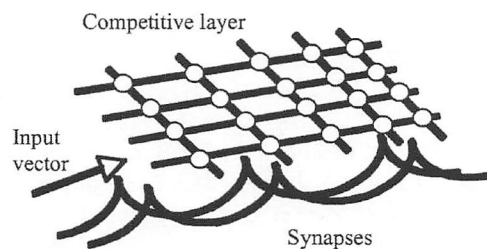


Fig. 1 SOM model

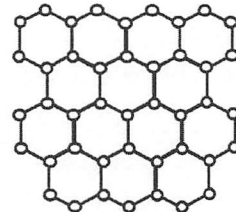


Fig. 2 Adopted topology (Hexagonal net)

入力ベクトル X は全てのニューロンに提示され, 2 次元トポロジーの頂点上の点 (i, j) にあるニューロン U_{ij} はその入力ベクトル X に結合した結合係数ベクトル W_{ij} を持っている.

3.2 学習アルゴリズム

以下に, SOM のアルゴリズムを示す.

- 1) 各結合係数 W_{ij} の初期値を乱数で与える.
- 2) 参照ベクトルからランダムに 1 要素 $V \in X$ を選択する.
- 3) 下式を満たす結合係数 W を式(1)により見つける.

$$\|V - W\| \leq \min_j \|V - W_{ij}\| \quad (1)$$

- 4) 式に従って, 結合係数 W の値を式(2)に従って更新する. ただし, $\alpha(t)$: 学習係数である.

$$W^{NEW} = W^{OLD} + \alpha(t)(V - W^{OLD}) \quad (2)$$

- 5) 結合係数 W を持つニューロン U の近傍にあるニューロンの結合係数も, 4) と同様の更新を行う. 近傍の数は近傍係数 $d(t)$ により決定する.
- 打ち切り条件を満たせば終了する. そうでなければ 2) に戻る

ここで, t は現在の学習回数とし, 学習係数 α は式(3)により, 学習回数によって減少する定数とする.

$$\alpha(t) = \frac{\alpha_0}{1 + e^{-A(BT-t)}} \quad (3)$$

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 $d(t)$ は式(4)により変更する。

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

4. SOM による分類法

4.1. SOM による分類

ウェブ上に公表されている、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から癌の予後決定に大きく関わる遺伝子を選定するため、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遺伝子を以下のように選定する。ここで、クラスとは 5 年後に生存したグループの患者遺伝子 (クラス 0) と 5 年以内に死亡したグループの患者遺伝子 (クラス 1) を示す。

58 人のもつ遺伝子 i の測定量 $G(i)$ ($i=1,2,\dots,7129$) から以下の 2 つの条件を満たすものを選定する。

$$|G(i)_{\max} - G(i)_{\min}| < 100 \text{ and } |G(i)_{\max} / G(i)_{\min}| < 3$$

ただし、ここで $G(i)_{\max}$ 及び $G(i)_{\min}$ は、遺伝子 $G(i)$ の測定量の最大値及び最小値である。

次に、上の条件を満たした遺伝子に対して、式(5)の評価 τ を適用し、上位 50 個をクラス 0 の遺伝子、下位 50 個をクラス 1 の遺伝子とした。

$$\tau_q = \frac{\mu_{q0} - \mu_{q1}}{\sigma_{q0} - \sigma_{q1}} \quad (5)$$

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。

4.2. データの正規化

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。

$$G(i) = \frac{G(i) - G(i)_{\min}}{G(i)_{\max} - G(i)_{\min}} \quad (6)$$

5. SOM による遺伝子の分類

5.1. SOM の適用

前処理を終えた各遺伝子を入力ベクトル \mathbf{X} 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 3 に示す。実験条件は表 1 の様に設定した。

Table 1 Experimental conditions

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 3 では、青三角形ラベルがクラス 0 の入力ベクトル、赤三角形ラベルがクラス 1 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、4 つの領域に患者群が分類されているのがわかる。

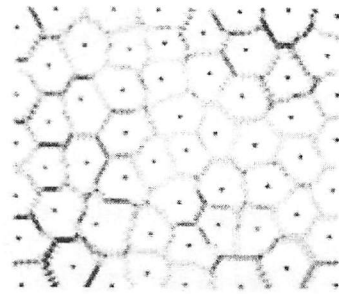


Fig. 3 Result of classified patients' genes

5.2. SOM による遺伝子系統図の作成

分類済みの図 3 の SOM マップを利用し、そこから最小結合木構造を遺伝子系統図として作成する。以下に系統図作成のためのアルゴリズムを示す

- 1) 参照ベクトル \mathbf{X}_i を 1 つ選択する
- 2) ベクトル \mathbf{X}_i から残りのベクトル \mathbf{X}_j への類似度を、それらのベクトルのなす $\cos\theta$ とし、以下から求める。

$$\cos\theta = \frac{\mathbf{X}_i^T \mathbf{X}_j}{\|\mathbf{X}_i\| \|\mathbf{X}_j\|} \quad (7)$$

- 3) 患者 i の遺伝子につける枝の最大分岐数を 4 個とする。
- 4) $0 \leq \cos\theta < 0.25$ の遺伝子が複数あれば、これらを分岐遺伝子として、枝を分岐する。4 以上の場合は、最大分岐を 4 とする。
- 5) 全ての遺伝子が $0.25 \leq \cos\theta < 1.0$ であれば、その中で最小の値をもつ遺伝子を遺伝子 i から伸長する遺伝子とする。
- 6) 患者 i の遺伝子データを取り除く。
- 7) 系統図の先端の患者の遺伝子 (グラフの葉) \mathbf{X}_i を選択する。
- 8) 全ての患者の遺伝子データが系統図に使用されたら終了する。そうでなければ、2)へ戻る。

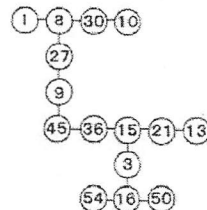


Fig. 4 A part of patients' gene tree

5.3. 樹枝図からの予後遺伝子選定

図 4 に結果の一部を示した。図 4 に示される系統図の中で、クラス 0 とクラス 1 との境界にある遺伝子間で、遺伝子データの各項目において差分解析を行い、癌の予後決定遺伝子を推測する。

6. おわりに

新たに遺伝子解析ツールとして系統図を自動生成する方法を SOM に基づいて提案した。今後、より詳細な系統図を作成し、予後遺伝子の特定を調べる予定である。

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)