

情報論的アプローチによる 新規治療標的分子の探索

東京理科大学 理工学部 情報科学科 准教授 ^{さとう}佐藤 ^{けいこ}圭子

東京理科大学 理工学部 情報科学科 教授 ^{たはた}田畑 ^{こうじ}耕治

はじめに

難治性がんの早期発見，そして分子標的治療の開発に向けての基盤を形成するために，多くの患者の臨床データ，および大量の遺伝子の発現データや変異データをどう扱い，どう解析すればよいのかを考えることは非常に重要です。また，たとえ同じがん種の患者間でも様々な異なる分子的特徴を示し，その病気の進行や治療効果は異なります。それゆえ，患者の臨床経過や治療効果の違いに関連する分子特性を様々な観点から明らかにすることは，臨床的に適用可能な分子ベースの分類システムや新しい標的治療の開発にも不可欠です。筆者らは，これまでに，乳がん患者や神経膠腫患者の的確な診断や予後分類のためのバイオマーカーを同定し，標的治療法の将来の開発のための基礎を築くことを目的として，それらの患者の遺伝子発現パターンの包括的な解析に情報論的手法や統計的手法を適用してきました。予後とは，患者が今後どのような薬に反応するのかとか，生存期間はどのくらいになるのかといった医学上の見通しのことです。例えば，この遺伝子のここにこういった変異があるとこの薬は効かないとか，この遺伝子が大量にタンパク質を作り出している場合は，生存期間は非常に短く2-3年だろうとか，がん診断時にこのように予後を判断できる遺伝子，すなわち，新しい治療薬のターゲットとなる可能性のある遺伝子を見つけ出す研究をしています。

情報論的手法

私たちは，現在，様々なデータベースから多くのデータを取得して使用することができます。がんのデータベースからも多くの患者の臨床データや遺伝子の発現データなどを手にすることができます。そして，発現データからは，患者個々の発現量，臨床データからは，診断時の年齢であったり，がんのステージであっ

たり，グレードであったり，生存期間であったり様々な情報を得ることができます。

私たちが，それらの情報を科学として扱うためには，数量化が必要となります。例えば，あるがんの再発に深く関わる特異的な発現を示す遺伝子を同定するのであれば，科学の領域では情報を量としてとらえる手法が確かに存在するので，再発情報や遺伝子の発現情報をきちんと数理的に用いることが必要です。

事象 x_i が確率 $p(x_i)$ で生起する完全事象系

$$X = \begin{pmatrix} x_1 & \cdots & x_m \\ p(x_1) & \cdots & p(x_m) \end{pmatrix}$$

のもつあいまいさは，

$$S(X) = -\sum_{i=1}^m p(x_i) \log p(x_i)$$

で表されます。これを系 X のエントロピーまたは情報量といいます。例えば，あるがん診断後5年間追跡調査された患者群において，再発状況（再発あり，再発なし）の事象からなる完全事象系を設定すると，再発に関する予後のあいまいさを量としてとらえることができます。

また，事象 y_j が確率 $p(y_j)$ で生起する完全事象系を

$$Y = \begin{pmatrix} y_1 & \cdots & y_n \\ p(y_1) & \cdots & p(y_n) \end{pmatrix}$$

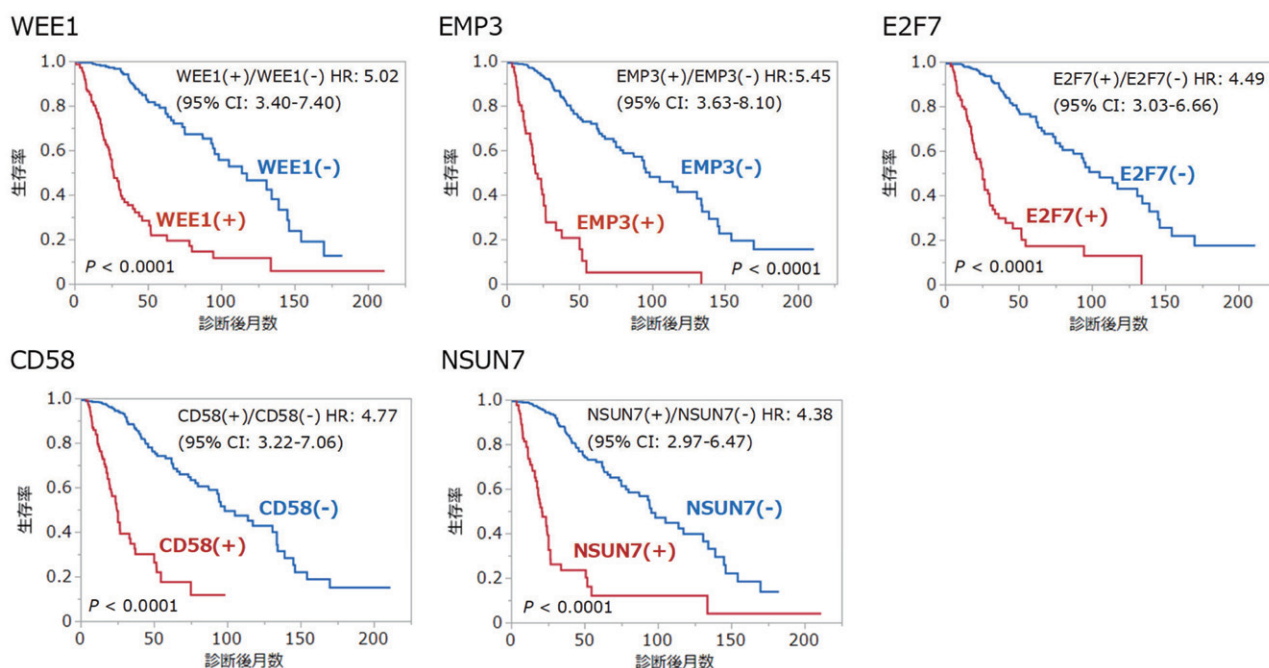
とすると， X と Y の複合完全事象系は，

$$X \times Y = \begin{pmatrix} x_1 y_1 & \cdots & x_m y_n \\ p(x_1, y_1) & \cdots & p(x_m, y_n) \end{pmatrix}$$

で表せます。なお， $p(x_i, y_j)$ は事象 $x_i \in X$ と事象 $y_j \in Y$ が同時に生起する確率です。このとき， Y を観測することによって得られる X の情報量は相互情報量と呼ばれ，

$$MI(X, Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

で与えられます。先の例を用いて， X を再発状況の事象からなる完全事象系， Y をある遺伝子の発現状況（過剰発現，過少発現）の事象からなる完全事象系，



【図 1】各遺伝子の発現量の高 (+) 低 (-) でグループ分けした LGG 患者群の生存曲線、および診断時の年齢と悪性度で調整されたハザード比 (HR)

$X \times Y$ をその複合事象からなる複合完全事象系として設定すると、相互情報量を計算することができます。これは、再発に関する予後のあいまいさが、遺伝子の発現量を観測することによってどれだけ減少したかを示す量としてとらえることができます。相互情報量が大きければ大きいほど X のあいまいさは削減されます。なお、相互情報量が 0 ならば 2 つの事象系 X と Y は相関がありません (独立です)。

神経膠腫 (グリオーマ) 患者の予後因子

ここで、上記の情報論的手法を取り入れたがんの研究の一つを紹介します¹⁾。悪性の脳腫瘍の一つである神経膠腫の悪性度 (グレード) 2, 3 に分類される lower grade glioma (LGG) の患者は、グレード 4 のグリオーマである生存期間が非常に短い glioblastoma (GBM) と呼ばれる患者と比べて長く生存します。筆者らが、527 名の LGG 患者と 164 名の GBM 患者の診断後の 5 年生存率を調べたところ、それぞれ、61.6% と 2.3% でした。なお、LGG 患者の診断時の年齢の中央値 (最小値, 最大値) は 41 歳 (14 歳, 87 歳) で、257 名がグレード 2 で 270 名がグレード 3 に分類されていました。GBM 患者の診断時の年齢の中央値 (最小値, 最大値) は 60 歳 (21 歳, 89 歳) でした。一般に、GBM 患者の平均生存期間は 1 年余りであるのに対し、LGG 患者の転帰は多様で、数か月で GBM に進行する患者もいれば、何年も進行しないま

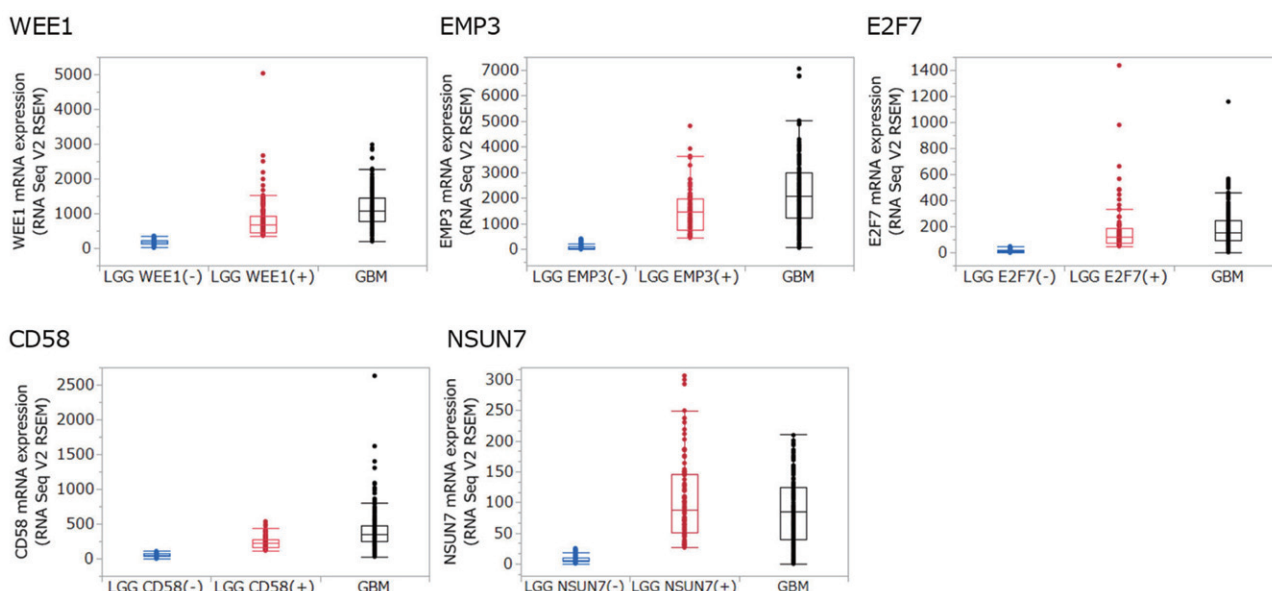
ま生存している患者もいます。筆者らは、情報論的アプローチから LGG 患者の進行や生存期間の違いに関与する予後因子を見つけ出すのに、先に述べた 527 名の LGG 患者の 18,413 に及ぶ遺伝子の発現データおよび生存期間のデータを使用しました。その結果、5 つの遺伝子 (WEE1, EMP3, E2F7, CD58, NSUN7) の高発現グループは低発現グループと比べ生存期間が著しく短く、これらの遺伝子が高発現である場合、LGG 患者の死亡リスクを高めることが分かりました【図 1】。筆者らは、LGG 患者の診断時の年齢やグレードによる影響を取り除いたとしても高発現群と低発現群の間で死亡のリスクにほぼ 4~5 倍の差がある 5 つの遺伝子を見つけ出すことができました。さらに、LGG における短期生存に関連するこれらの 5 つの遺伝子の高発現のパターンは、先に述べたグレード 4 の 164 名の GBM 患者でも確認できました【図 2】。

分割表解析

完全事象系 X と Y , 複合完全事象系 $X \times Y$ を表の形式で表現すると以下のようにまとめられる【表 1】。

【表 1】同時確率分布

X	Y			合計
	y_1	$\cdots$	y_n	
x_1	$p(x_1, y_1)$	$\cdots$	$p(x_1, y_n)$	$p(x_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	$p(x_m, y_1)$	$\cdots$	$p(x_m, y_n)$	$p(x_m)$
合計	$p(y_1)$	$\cdots$	$p(y_n)$	1



【図2】各遺伝子の発現量の高（+）低（-）でグループ分けしたLGG患者群とGBM患者群の発現量の比較（箱ひげ図およびストリッププロット）

一方で、変数 X と Y について調査・実験などから得られたデータをクロス集計した表を以下のように表すことにする【表2】。

【表2】観測度数

X	Y			合計
	y_1	$\cdots$	y_n	
x_1	$n(x_1, y_1)$	$\cdots$	$n(x_1, y_n)$	$n(x_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	$n(x_m, y_1)$	$\cdots$	$n(x_m, y_n)$	$n(x_m)$
合計	$n(y_1)$	$\cdots$	$n(y_n)$	n

この表のことを統計学では、分割表と呼ぶ。私たちは、得られた分割表が、上で与えた未知の確率分布（表1）に基づいて生じたと考えることにより、分割表から同時確率分布 $p(x_i, y_j)$ を推測する。 $p(x_i, y_j)$ を推測するために、 $n(x_i, y_j)/n$ がよく用いられる。

分割表解析において興味のある仮説は、「変数 X と Y は互いに統計的独立」である。この仮説は、

$$p(x_i, y_j) = p(x_i)p(y_j) \quad (i=1, \dots, m; j=1, \dots, n)$$

と表せる。この仮説が成り立つ場合には、事象 y_j が起こったという条件の下で事象 x_i が起こる条件付き確率 $p(x_i|y_j)$ は、

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p(x_i)p(y_j)}{p(y_j)} = p(x_i)$$

となる。つまり、変数 X と Y が互いに統計的独立である場合には、条件付き確率と無条件での確率が同じという意味で、変数 Y は変数 X に対して情報を与えないことが確認できる。変数 X を再発状況（再発あり、再発なし）、変数 Y をある遺伝子の発現状況（過剰発現、過少発現）としたとき、変数 X と Y が互いに統計的

独立であれば、遺伝子の発現状況は、再発状況に情報を与えないと解釈することができる。逆に、変数 X と Y が互いに統計的独立でない場合には、遺伝子の発現状況は、再発状況に関する情報を与える可能性を示唆する。

この統計的独立性の仮説が得られた分割表データに適合しているかどうかを調べる方法として、独立性のための Pearson のカイ二乗検定がよく用いられる。特に、観測度数の合計 n が小さい 2×2 分割表の場合には、Fisher の正確検定も知られている。

情報論的手法と分割表解析

得られた分割表データに対して、独立性のためのカイ二乗検定によって仮説「変数 X と Y が互いに統計的独立である」が棄却されたとする（得られた分割表データに統計的独立性が適合していない）。このとき、私たちの次なる関心は変数 X と Y の連関の度合いを測ることである。連関の度合いについては、Pearson の積率相関係数を思い出すと良い。連続的な変数 x と y に関するペアデータ (x_i, y_i) を二次元平面にプロットしたものを散布図というが、散布図の様子が全体的に右肩上がり、つまり変数 x の値が増加すると変数 y の値も増加する傾向を正の相関、逆に、散布図の様子が全体的に右肩下がり、つまり変数 x の値が増加すると変数 y の値が減少する傾向を負の相関と呼んだ。相関係数は、 -1 以上 1 以下の値を取り、相関係数が 1 のときを正の完全相関、相関係数が -1 のときを負の完全相関という。この積率相関係数に対応するよう

な変数 X と Y の連関を表現する尺度として, Proportional Reduction in Variation (PRV) 尺度が次のように与えられる.

$$\frac{V(Y) - E[V(Y|X)]}{V(Y)}$$

ただし, $V(Y)$ は変数 Y に対する variation であり, $E[V(Y|X)]$ は変数 X が与えられたもとでの変数 Y の条件付き variation の変数 X についての期待値である. この PRV 尺度の特別な場合として, Goodman and Kruskal の τ , Theil の uncertainty coefficient がある.

PRV 尺度と相互情報量の関係について説明する. variation として, Shannon エントロピー

$$V(Y) = - \sum_{j=1}^n p(y_j) \log p(y_j)$$

を用いる. このとき, $X=x_i$ が与えられたもとでの変数 Y の条件付き variation は,

$$V(Y|x_i) = - \sum_{j=1}^n p(y_j|x_i) \log p(y_j|x_i)$$

で与えられる. このことから,

$$V(Y) - E[V(Y|X)] = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

が得られる. つまり, $V(Y) - E[V(Y|X)] = MI(X, Y)$ であり, PRV 尺度は相互情報量 $MI(X, Y)$ を Shannon エントロピー $S(Y)$ で正規化したものと考えられる. これは, Theil の uncertainty coefficient である.

筆者らは, グリオーマ患者の予後因子に関する研究において, 情報論的アプローチによって 5 つの遺伝子をピックアップすることに成功した. このことから, 相互情報量と関連のある PRV 尺度を用いた方法によって新たな関連遺伝子を発見することが可能かもしれない. そこで, 筆者らは富澤貞男先生 (明星大学, 東京理科大学名誉教授) が開発した一般化 PRV 尺度に注目した²⁾.

一般化 PRV 尺度について紹介する. 一般化 PRV 尺度は, Patil and Taillie の diversity index of degree- λ を variation として採用したものである. つまり,

$$V(Y) = \frac{1}{\lambda} \left(1 - \sum_{j=1}^n p(y_j)^{\lambda+1} \right) \quad (\lambda > -1)$$

であり, $X=x_i$ が与えられたもとでの変数 Y の条件付き variation は,

$$V(Y|x_i) = \frac{1}{\lambda} \left(1 - \sum_{j=1}^n \left(\frac{p(x_i, y_j)}{p(x_i)} \right)^{\lambda+1} \right)$$

で与えられる. このことから, 一般化 PRV 尺度は

$$\frac{\sum_{i,j} \frac{p(x_i, y_j)^{\lambda+1}}{p(x_i)^{\lambda}} - \sum_{j=1}^n p(y_j)^{\lambda+1}}{1 - \sum_{j=1}^n p(y_j)^{\lambda+1}} \quad (\lambda > -1)$$

で定義される. ただし, $\lambda=0$ においては $\lambda \rightarrow 0$ の極限で定義され, Theil の uncertainty coefficient に一致する. 一般化 PRV 尺度は, 0 以上 1 以下の値をとり, 尺度の値が 0 であるための必要十分条件は変数 X と Y が互いに統計的独立であることであり, 尺度の値が 1 であるための必要十分条件は条件付き variation がないとき (つまり, 各 x_i に対して, $p(x_i, y_j)/p(x_i)=1$ を満たすある y_j が存在する) である. このことから, 一般化 PRV 尺度の値が 1 に近ければ近いほど, 変数 X と Y の連関は強いと考えることができる. 言い換えれば, 変数 X を知ることによって, 変数 Y をよりよく予測できることを示している.

おわりに

本稿では, (1) 情報論的アプローチ, (2) 情報論的アプローチによる新規治療標的分子探索の成功例, (3) 分割表解析, (4) 情報論的アプローチと分割表解析の関係, の 4 つについて簡単に紹介した.

現在, 筆者らは, 一般化 PRV 尺度を用いて, 様々ながんに対する新規治療標的分子探索を実施している. いくつかの遺伝子をピックアップすることに成功しているが, 既存の手法との比較, 情報論的アプローチ特有の性質の検討, 打ち切りデータへの対応方法などクリアしなければならない課題が多く存在する. これらの課題をクリアし, 更なる情報論的アプローチの開発・研究を進め, 難治性がんの早期発見, そして分子標的治療の開発に向けての基盤形成に貢献したい.

【参考文献】

- 1) Sato, K., et al. (2020). Five genes associated with survival in patients with lower-grade gliomas were identified by information-theoretical analysis. *Anticancer Research*, 40, 2777-2785.
- 2) Tomizawa, S., et al. (1997). Generalized proportional reduction in variation measure for two-way contingency tables. *Behaviormetrika*, 24, 193-201.

