Cluster Validity Indexを利用した遺伝子発現差解析手法の評価

Cluster Validity Indexを利用した遺伝子発現差解析手法の評価

Cluster Validity Indexを利用した遺伝子発現差解析手法の評価

小野　修司[1)]，澤井　政宏[2)]，田中　秀典[3)]，長島　知正[4)]，岡田　吉史[5)]
1）室蘭工業大学情報工学専攻（M2）
2）室蘭工業大学工学研究科生産情報システム工学専攻（D2）
3）室蘭工業大学SVBL　4）室蘭工業大学情報工学科　5）産業技術総合研究所

1．序論

近年、Affymetrix社のGeneChipの登場により、遺伝子発現量を網羅的に捉えることが可能となった。しかし、GeneChipからもたらされる遺伝子発現データの量は膨大なため、計算機での処理が必須となっている。GeneChipの利用方法の1つに遺伝子発現差解析がある。遺伝子発現差解析は、例えば、健常者群と患者群間の遺伝子発現量の差を統計的に見積もることで、疾患遺伝子として尤もらしい度合いを表すスコアを遺伝子ごとに算出する。解析者は、このスコアを基に疾患遺伝子の同定を行う。遺伝子発現差解析の手法はいくつか提案されている[1]が、どの手法も利点と欠点を併せ持っている。そのため、遺伝子発現データセットの傾向によって、最適な手法は異なる。

疾患遺伝子はその疾患に対して特徴的に発現する遺伝子であるため、その疾患遺伝子を用いて患者のクラスタリング[2]を行った場合、各疾患を持つ患者ごとのクラスターを生成され、かつ各クラスター間の距離が大きくなると考えられる。よって各遺伝子発現差解析手法によって疾患遺伝子の可能性が高いと判断された遺伝子を用いて患者のクラスタリングを行い、得られたクラスターの妥当性を評価することによって、どの手法が最適かを判断することが可能であると考えられる。

そこで本研究では、急性リンパ性白血病(ALL)と急性骨髄性白血病(AML)、嚢胞性繊維症(CF)と健常者の二つの遺伝子発現データセット[3]について、t検定、相関比、SAMの三手法によって遺伝子発現差解析を行い、どのような傾向を持つ遺伝子発現データセットにどの手法が最適であるかを、クラスターの妥当性を表すCluster Validity Index[4]を用いて評価する。

2．遺伝子発現差解析

遺伝子発現差解析の目的は遺伝子の発現量の差から何らかの特徴的な意味を見出すことであるが、この解析手法は多種多様である。

現在、遺伝子発現差解析手法の多くは二群間の相関の強さを知ることのできる統計的検定手法や分散分析手法に基づいて行われている。本研究では一般的に用いられているt検定、相関比、SAMの三つの手法を対象とした。

表1．遺伝子発現データセットの例

	AML患者				ALL患者			
遺伝子名	患者1	患者2	・・・	患者24	患者1	患者2	・・・	患者28
遺伝子A	404.4013	497.9097	・・・	854.6865	850.048	1252.852	・・・	1118.824
遺伝子B	1166.468	780.7663	・・・	818.2772	1314.173	1319.951	・・・	1199.308
遺伝子C	316.2657	196.8496	・・・	511.5075	103.7123	296.1054	・・・	327.6257
遺伝子D	59.04235	17.49949	・・・	11.4694	44.93436	48.7953	・・・	27.79331
遺伝子E	150.3723	71.49325	・・・	65.39445	132.0143	165.846	・・・	117.0784
遺伝子F	125.0104	140.8753	・・・	294.2496	371.3474	313.5573	・・・	280.6461
遺伝子G	5625.491	10512.55	・・・	3042.252	7106.59	15767.81	・・・	4474.031
遺伝子H	20.90229	23.28805	・・・	9.848327	13.32856	172.4844	・・・	15.05943
遺伝子I	935.7613	475.9129	・・・	809.6168	492.684	594.054	・・・	486.2627
遺伝子J	2569.851	2580.148	・・・	3287.649	5682.981	5212.962	・・・	3191.398
遺伝子K	2930.512	10884.5	・・・	4122.077	8133.776	2769.905	・・・	4551.036
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3．GeneChip

遺伝子発現差解析を行うためには、疾患を持つ患者の遺伝子発現量データを得る必要がある。GeneChipから数値化された遺伝子発現量データを得る過程は以下の4つの手順に分けられる。

①：ターゲットRNAから順にcRNAを合成し、その際に標識物質で標識する。

②：標識されたcRNAを断片化し、GeneChipのプローブにハイブリダイズする。

③：GeneChipを洗浄し、ターゲットを蛍光標識する。

④：蛍光標識されたGeneChipをスキャニングすることで、イメージ画像(遺伝子発現データ)を取得し、各遺伝子の輝度を数値化する。この処理を正規化と呼ぶ。

なお、正規化にはGeneChip付属の画像処理ソフトウェアであるMAS5(Affymetrix Microarray Suite 5.0)を使用する。数値化した遺伝子発現データセットの例を表1に示す。

4．手法

4.1　遺伝子発現差解析手法

GeneChipから得られた遺伝子発現量に、MAS5による正規化を行う。t検定、相関比、SAMの三つの手法によって遺伝子発現差解析を行い、疾患遺伝子として尤もらしい度合いを表すスコアを遺伝子ごとに算出する。その後、遺伝子をスコアに基づいてランキングする。

4.1.1　t検定

帰無仮説が正しいと仮定した場合に、統計量t_0がt分布に従うことを利用する統計学的検定法。二群の平均値の差について検定で、P値が小さいほど、「平均に差がある」確率が高くなり有意差があるといえる。本研究では、welchのt検定を用いた。

$$t_0 = \frac{\left|\overline{X_a} - \overline{X_b}\right|}{\sqrt{U_a/n_a + U_b/n_b}}$$

$$\nu = \frac{\left(U_a/n_a + U_b/n_b\right)^2}{\left(U_a/n_a\right)^2/\left(n_a - 1\right) + \left(U_b/n_b\right)^2/\left(n_b - 1\right)}$$

$$P = \Pr\left\{|t| \geq t_0\right\}$$

$\overline{X}_a$, $\overline{X}_b$ ：疾患A,Bの発現量の平均値

n_a, n_b ：疾患A,Bの患者数　ν ：自由度

U_a, U_b ：疾患A,Bの不偏分散

4.1.2 相関比

二群間の関連を示す指標で、相関比の値が大きいほど、その遺伝子が 2 つの疾患を良く区別する遺伝子であることを示す。相関係数が線形の分布にしか対応できないのに対し、分布が非線形でも対応可能である。相関比の値は0から1の間を取る。

全体分散　$\sigma^2 = \sum_{j}^{2}\sum_{i}^{nj}\left(x_i(j) - \tilde{x}\right)^2$

群間分散　$\sigma_B^2 = \sum_{j}^{2} n_j\left(\tilde{x}_j - \tilde{x}\right)^2$

相関比　$\eta^2 = \frac{\sigma_B^2}{\sigma^2}$

n_j ：疾患 j の患者数

$\tilde{x}$ ：発現量の平均　$\tilde{x}_j$ ：疾患 j の発現量の平均

$x_i(j)$ ：疾患 j の患者 i の発現量

4.1.3 SAM

スタンフォード大学が開発したソフトウェア。DNAマイクロアレイデータにおいて、なんらかの性質と有意に相関する遺伝子を検出することを主目的とする。$d(i)$の値の大きさが相関の強さを表す。

$$d(i) = \frac{\overline{x_I}(i) - \overline{x_U}(i)}{s(i) + s_0}$$

$\overline{x}_I(i)$, $\overline{x}_U(i)$ ：疾患 I, Uの遺伝子 i の発現量の平均

$s(i)$ ：遺伝子 i の標準偏差　s_0 ：fudge factor

4.2 各遺伝子発現差解析手法の評価

4．1節に示した手法によってランキングされた上位 n 個の遺伝子（以下、上位 n 遺伝子）を用いて k-means によって遺伝子発現データセットのクラスタリングを行う。ここで、nは2～各データセットに含まれる遺伝子数まで変化させるものとする。そのとき、各クラスタリング結果に対して Cluster Validity Index を求めグラフを作成する。このグラフを比較検討することによりどの手法がどのような遺伝子発現データセットに対して適切かを評価する。

4.2.1 k-means

非階層的クラスタリングの一つで、データをその類似性に応じてk個のクラスターに分類する手法である[5]。データを多次元空間上の点と見なし、距離の近いものを同一のクラスターに分類する。本研究では患者 1 検体を 1 つのデータとし、各データを 2 つのクラスターに分類する。データは、各患者における上位n遺伝子の発現量から構成される、n次元のベクトルデータである。

4.2.2 Cluster Validity Index

クラスターの妥当性を定量的に評価するための指標で、統計的に最適なクラスター数を見積もる場合などに用いられる。本研究では遺伝子発現差解析手法の評価に用いるため、

1）　どれだけ正確にクラスタリングが行われているか

2）　どれだけクラスター間の重心の距離を遠くできているか

という二つの観点が重要となる。そこで、Cluster Validity Index の一部である、Cluster Entropy(CE) と Cluster Separation(CS)の二つの指標を使用し、評価を行う。

・Cluster Entropy

クラスターの乱雑さの度合いを評価するための指標。値が 0 に近づくほど乱雑さのない正確なクラスタリング結果であることを表す。*Entropy*の算出にはクラスター毎の *Entropy* を求める必要がある。

$$Entropy_i = -\sum_j \frac{n(l_j, c_i)}{n(c_i)} \log \frac{n(l_j, c_i)}{n(c_i)}$$

$$Entropy = \frac{1}{\sum_i n(c_i)} \sum_i n(c_i) Entropy_i$$

l_j ：本来のクラスター j

c_i ：クラスタリングによって生成されたクラスター i

$n(c_i)$ ：クラスター i の患者数

$n(l_j, c_i)$ ：クラスター i の中の l_j の数

・Cluster Separation

クラスター間の重心の距離を表す指標。値が小さいほどクラスター間の重心の距離が遠いことを表す。

$$Separation = \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{i=1, j\neq i}^{c} \exp\left(-\frac{d^2\left(x_{c_i}, x_{c_j}\right)}{2\sigma^2}\right)$$

c：クラスター数　c_i：クラスターi　c_j：クラスターj
σ：ガウス定数　$d^2\left(x_{c_i}, x_{c_j}\right)$：$c_i$と$c_j$の重心の距離

本研究では、c=2 とする。また、ガウス定数σは通常σ=1 として用いられるが、本研究ではデータの性質上σ=100 として計算する。

5　実験

5.1　データ

ALL24 検体と AML28 検体の遺伝子発現データセット（以下、ALL-AML セット）と CF9 検体と健常者 9 検体の遺伝子発現データセット（以下、CF セット）二つを用いて、４章で示した手法によってどのような傾向を持つ遺伝子発現データセットにどの遺伝子発現差解析手法が最適であるかを評価する。

5.2　結果

結果を図１−８に示す。図１は ALL-AML セットの CE、図２は ALL-AML セットの CS、図３は CF セットの CE、図４は CF セットの CS である。また、遺伝子発現差解析では上位にランキングされた遺伝子ほど重要であるため、立ち上がり上位 400 位までの結果をそれぞれ図５、図６、図７、図８に示す。

図１の ALL-AML セットの CE は遺伝子の数が増えるにつれて値が上昇する。それに対し、図３の CF セットでは、ほぼ水平に推移している。CE の値が 0 の時、クラスターは完全に分割されている状態であり、ALL-AML セットはどの手法も遺伝子数約 3000 まで正確なクラスタリングが可能であることがわかる。図２、４の CS のグラフでは、ALL-AML セットと CF セットどちらの結果でも遺伝子数が増加するに従い CS の値は減少している。

図５、７の立ち上がりのグラフでは、CE の値はどちらのデータに対しても相関比が最もよい値を示していた。また、図６、８の CS の値は SAM がよい値を示した。

5.3　考察

図１、３に示されるとおり、ALL-AML セットの CE は低い値を取っており、CF セットの CE は高い値を取っている。また、各遺伝子発現差解析のスコアを確認したところ、ALL-AML セットは高いスコアを持つ（t 検定においては低いスコアを持つ）遺伝子の数が多く、CF セットは高いスコアを持つ遺伝子の数が少なかった。これはすなわち、ALL-AML セットは各疾患に対して特徴的に発現する遺伝子数が多いため 2 つの群を容易に分類可能なデータであり、CF セットは各疾患に対して特徴的に発現する遺伝子数が少ないため 2 つの群を分類することが困難なデータであることを示している。

図５−８より、k-means を使用した場合、最も正確なクラスターを生成する遺伝子を上位にランキングできたのは相関比であり、クラスター間の距離を広げられる遺伝子を上位にランキングしたのは SAM であった。

遺伝子発現差解析の手法として、正確なクラスターを生成することは必要条件で、これを満たしておりかつクラスター間の距離を広げられるものが適切な手法である。このことから、いずれかの手法を用いて遺伝子発現差解析を行った結果、ALL-AML セットのように各疾患に対して特徴的に発現する遺伝子（スコアの高い遺伝子）が多数見つけられるデータセットには CS を小さくすることに適した SAM を使用することが適切であると考えられる。逆に CF セットのような各疾患に対して特徴的に発現する遺伝子の数が少ない場合には CE を小さくすることが可能な相関比を用いるのが適切であると考えられる。

6．　総括

本研究では、どのような傾向を持つ遺伝子発現データセットにどのような遺伝子発現差解析手法が適切かを明らかにすることを目的に、実験を行った。実験結果より、各疾患に対して特徴的に発現する遺伝子が多数見つけられる遺伝子発現データセットに対しては SAM、そうでないものには相関比を使用することが望ましいことが明らかになった。これにより、生物学者が疾患遺伝子を同定する際の能率の向上が期待される。

今後の課題として、別の遺伝子発現データセットを用いることで結果の再現性を調べること、他の遺伝子発現差解析手法について同様の実験を行うことが必要である。

参考文献

[1] V. G. Tusher, R. Tibshirani, and G. Chu : Significance analysis of microarrays applied to the ionizing radiation response, PNAS, Vol. 98 no. 9, pp. 5116-5121, 2001

[2] A. k. Jain, M. N. Murty, P. J. Flynn : Data clustering a Review. ACM. Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999

[3] BROAD INSTITUTE : http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63

[4] J. He, A. Tan, and C. Tan : Modified ART 2A Growing Network Capable of Generating a Fixed Number of Nodes. IEEE TRANSACTIONS ON NEURAL NETWORKS, Vol. 15, NO. 3, pp. 728-737, 2004

[5] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, Church GM. : Systematic determination of genetic network architecture, Nat. Genet., pp. 281-285, 1999

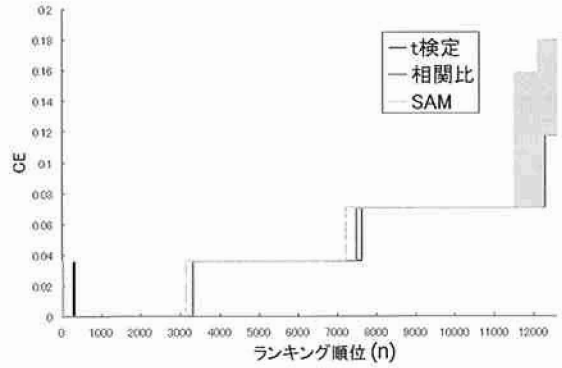

図1．ALL-AMLセットにおけるCluster Entropy

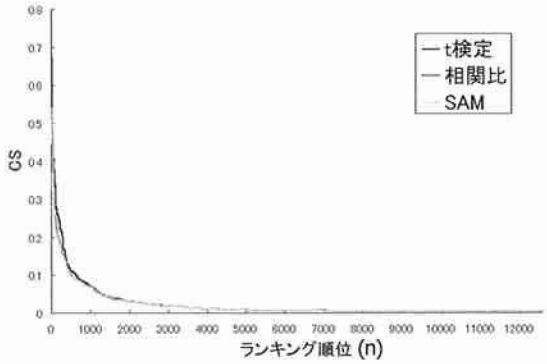

図2．ALL-AMLセットにおけるCluster Separation

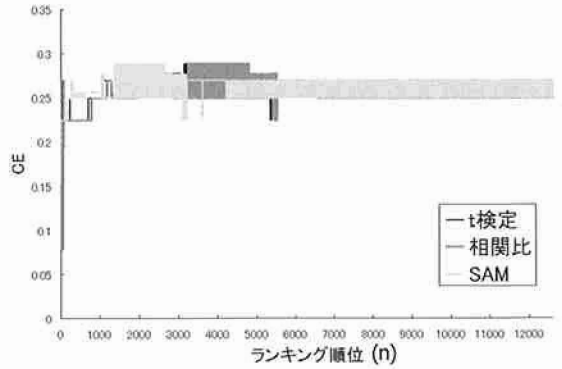

図3．CFセットにおけるCluster Entropy

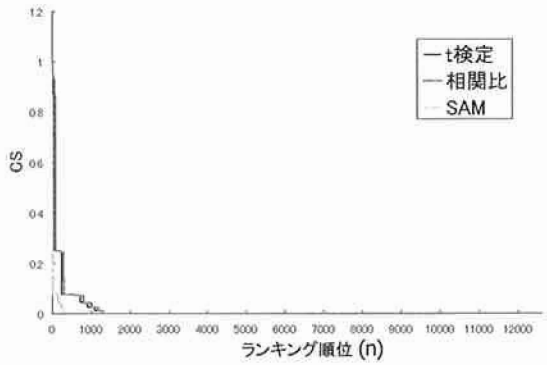

図4．CFセットにおけるCluster Separation

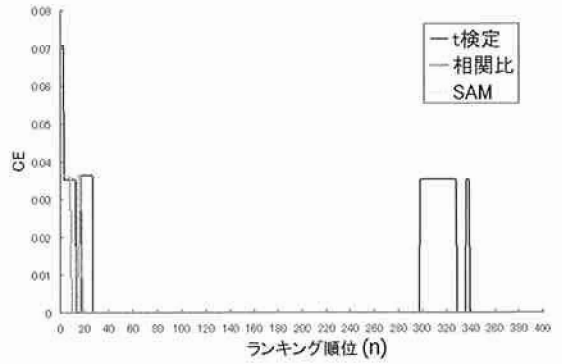

図5．ALL-AMLセットにおけるCluster Entropyの立ち上がり

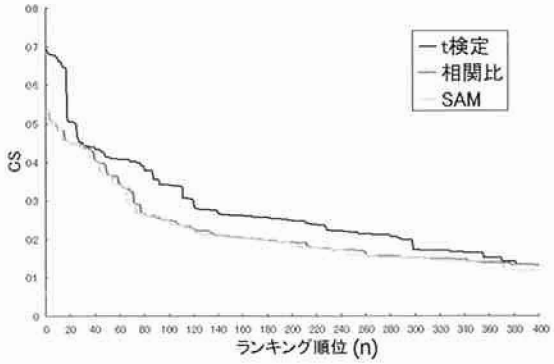

図6．ALL-AMLセットにおけるCluster Separationの立ち上がり

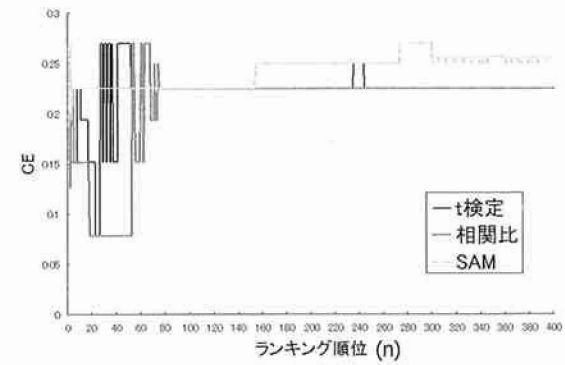

図7．CFセットにおけるCluster Entropyの立ち上がり

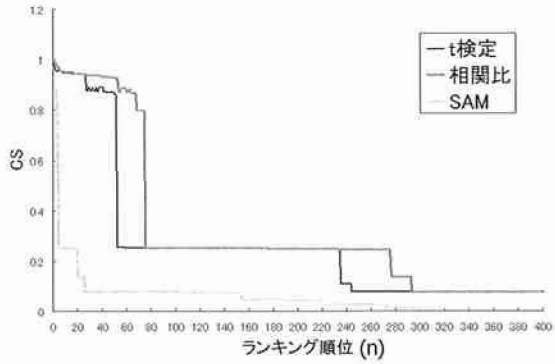

図8．CFセットにおけるCluster Separationの立ち上がり