DNA microarray データを検定しやすく変換する
Transformation of DNA microarray data for statistical tests

井上 真郷 (PY)[1,4], 西村 信一[1,5], 堀 玄[2], 甘利 俊一[1], 斉藤 美知子[3], 吉原 良浩[3], 中原 裕之[1]

Masato Inoue[1,4], Shin-ichi Nishimura[1,5], Gen Hori[2], Shun-ichi Amari[1],

Michiko Saito[3], Yoshihiro Yoshihara[3], Hiroyuki Nakahara[1]

理化学研究所 脳科学総合研究センター
{[1]脳数理研究チーム, [2]脳信号処理研究チーム, [3]シナプス分子機構研究チーム},
[4]京都大学大学院医学研究科 耳鼻咽喉科・頭頸部外科, [5]東京大学大学院医学系研究科 耳鼻咽喉科学専攻
{[1]Lab. for Mathematical Neuroscience, [2]Lab. for Advanced Brain Signal Processing,
[3]Lab. for Neurobiology of Synapse}, RIKEN Brain Science Institute,
[4]Dept. of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto Univ.,
[5]Dept. of Otorhinolaryngology, and Head and Neck Surgery, Graduate School of Medicine, Univ. of Tokyo

E-mail: minoue@brain.riken.go.jp

Abstract

Biological data are often log-transformed to improve their normality and uniformity of variance. In DNA microarray data, however, the logarithmic transformation would fail because of non-positive data and nonuniformity of coefficients of variation. We developed a new logarithm-like transformation and validated it on a public dataset.

1 はじめに

遺伝子は，DNA 配列のみならず，その機能を発見する事が特許にもつながり，重要である．例えばヒトの遺伝子は 3〜4 万個と言われ，そのうち約 1 万個の機能が判明しているにすぎず，機能未知の遺伝子が，どのような組織でどのようなタイミングで発現しているかをスクリーニング検査できれば，機能を解明するのに大きな助けとなる．DNA microarray は，任意の生体組織について，数千〜数万個の予め選択した遺伝子がどの程度発現しているかを同時に調べる事ができる画期的な測定システムであり，これの登場により遺伝子研究が活性化する一方，大量のデータから統計的により正確な推測をする必要性が高まっている．

典型的な課題は，複数条件下で測定を繰り返し，条件間で発現量に差がある遺伝子を有意度順に並べるというものである．条件は，大脳と小脳といった組織別や，胎生 15 日目と 20 日目といった時系列，異なった量の薬物を投与するといった処置別など，何でもよい．今，1 万個の遺伝子について，各条件下で数回測定を繰り返したデータがあるとすると，条件間で発現の変化があるかどうかは分散分析を行えばよいし，薬物量との関係を調べるには回帰分析が使える．

しかし，これらの検定や解析は一般に「正規性」と「等分散性」を前提とする．即ち，各条件下での測定データは正規分布に従い，どの条件下でも共通の母分散を持っていなければならない．これが各遺伝子についてそれぞれ成立する事が求められる．分散分析はこれらの前提が少々崩れていても頑健性があり，また，non-parametric 検定という選択肢もあるが，この二つの性質は，他の多くの解析手法でも陰に陽に前提としており，分散分析に限らず重要である．我々は，主に oligonucleotide array と呼ばれるタイプの DNA microarray のデータに対して，この等分散性を改善する変換を考案したので報告する．

キーワード: DNA microarray，対数変換，変数変換，等分散性，正規性

2 モデル

2-1 対数正規分布

一般に，生物学的データは対数をとると正規分布に近くなる事が多い（対数正規分布）．即ち，観測値を X とすると，

$$X = e^{Z},\ Z \sim \mathrm{N}\left(\mu, \sigma^2\right) \tag{1}$$

となる．X の平均と分散が，

$$\mathrm{E}[X] = e^{\mu+\frac{\sigma^2}{2}},\ \mathrm{Var}[X] = \left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2} \tag{2}$$

となる事から，σ^2 が決まれば μ に関係なく，X の

分散が平均の二乗で表せる事，X の変動係数（標準偏差／平均の絶対値）が一定になる事が分かる．

$$\mathrm{Var}[X]=\left(e^{\sigma^2}-1\right)\{\mathrm{E}[X]\}^2$$
$$\sqrt{\mathrm{Var}[X]}/|\mathrm{E}[X]|=\sqrt{e^{\sigma^2}-1} \tag{3}$$

この事から逆に，得られたデータ（複数標本）について標本平均と不偏分散をプロットした時に分散が平均の二乗に比例する傾向があったり，変動係数が一定という傾向がある場合は，これらは共通の σ^2 を持つ対数正規分布である事が示唆され，対数変換を行う事で分散一定の正規分布に近い分布が得られる事が多い．

2-2 Oligonucleotide array

しかし，DNA microarray データ，特に oligonucleotide array (GeneChip®)データは，大まかには上記対数正規分布に近い性質を示すものの，1)非正値が含まれる，2)変動係数が平均が 0 付近で増大する，という癖があり，単純な対数変換では不十分である．

具体的には，測定される遺伝子発現量は発現時と非発現時で桁違いの値をとり，多く発現している時は 10000 程度で変動係数が 10～20%，あまり発現していない時は 10 程度で変動係数が 100%以上といった傾向があり，負のデータも 2～4 割程度混じっている．非正値があるため全てを対数変換できず，また，正値のみを変換しても（変動係数が一定でないため）分散が一定にならない．

Oligonucleotide array は，一つの遺伝子に対してその DNA 配列から M（≈20）ヶ所の部位を選んで別々に検出を行い，これらを平均して発現量 X とするので[1],

$$X=\frac{1}{M}\left(\sum_{m=1}^{M}Y_m+Q\right)$$
$$Y_m=e^{Z_m},\ Z_m\sim\mathrm{N}\left(\mu_Z,\sigma_Z^2\right),\ i.i.d. \tag{4}$$

と表す事ができる．Q は cross-hybridization 等のノイズである．この場合も，分散は平均の二次式で表せる．

$$\mathrm{Var}[X]=a\left(\mathrm{E}[X]-b\right)^2+c$$
$$a=\frac{1}{M}\left(e^{\sigma_Z^2}-1\right),\ b=\frac{1}{M}\mathrm{E}[Q],\ c=\frac{1}{M^2}\mathrm{Var}[Q] \tag{5}$$

後述するように，平均から分散が推定できれば，分散を一定に変換でき，等分散性を改善する事ができる．X は特定の遺伝子の特定の条件での発現量を考えているので，パラメータ a, b, c も遺伝子及び条件依存であるが，全データに対して共通の変換式を得るため，これらは共通とする（σ_Z^2，Q の分布，M を全遺伝子，全条件で同一とし，μ_Z のみ異なると仮定する）．また，X は M 個の平均なので，中心極限定理から正規分布と仮定する．

3 アルゴリズム

3-1 分散推定関数

ここでは，平均から分散を推定する関数を求める．改めて，遺伝子 i の条件 j での発現値を n_{ij} 回繰返し測定した値を $X_{ij}^{(k)}$ (k=1,2,...,n_{ij})と表記し，X_{ij} を確率変数として，平均及び分散を以下のようにに定義する．

$$\mu_{ij}=\mathrm{E}[X_{ij}],\quad \sigma_{ij}^2=\mathrm{Var}[X_{ij}] \tag{6}$$

X_{ij} は正規分布と仮定したので，標本平均，不偏分散はそれぞれ独立に正規分布，ガンマ分布に従う．

$$\overline{X}_{ij}=\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}X_{ij}^{(k)}\sim\mathrm{N}\left(\mu_{ij},\frac{\sigma_{ij}^2}{n_{ij}}\right)$$
$$S_{ij}^2=\frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}}\left(X_{ij}^{(k)}-\overline{X}_{ij}\right)^2 \tag{7}$$
$$\sim\mathrm{Ga}\left(\frac{n_{ij}-1}{2},\frac{2\sigma_{ij}^2}{n_{ij}-1}\right)$$

これらの尤度関数は，

$$\mathbf{M}=\left\{\mu_{ij}\right\},\ \mathbf{\Sigma}=\left\{\sigma_{ij}^2\right\},\ \overline{\mathbf{X}}=\left\{\overline{X}_{ij}\right\},\ \mathbf{S}=\left\{S_{ij}^2\right\} \tag{8}$$

として，

$$L\left(\mathbf{M},\mathbf{\Sigma};\ \overline{\mathbf{X}},\mathbf{S}\right)=$$
$$\prod_{i,j}f_{\mathrm{N}}\left(\overline{X}_{ij};\mu_{ij},\frac{\sigma_{ij}^2}{n_{ij}}\right)f_{\mathrm{Ga}}\left(S_{ij}^2;\frac{n_{ij}-1}{2},\frac{2\sigma_{ij}^2}{n_{ij}-1}\right) \tag{9}$$

と表現できる．但し，f_{N} と f_{Ga} は正規分布とガンマ分布の密度関数である．

$$f_{\mathrm{N}}\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$f_{\mathrm{Ga}}\left(y;v,\alpha\right)=\frac{1}{\alpha^v\Gamma(v)}y^{v-1}e^{-\frac{y}{\alpha}} \tag{10}$$

モデルより，

$$\sigma_{ij}^2 = a\left(\mu_{ij} - b\right)^2 + c \tag{11}$$

が成立するので，$\boldsymbol{\Sigma}$ をこれで置き換え，a, b, c を最尤推定で求める．

$$\arg\max_{a,b,c,\mathbf{M}} L\left(a,b,c,\mathbf{M};\ \overline{\mathbf{X}},\mathbf{S}\right) \tag{12}$$

これは，以下のように簡略化できる．

$$\arg\max_{a,b,c} \sum_{i,j} \max_{\mu_{ij}} n_{ij} \times \left[-\frac{V_{ij}^2 + \left(\overline{X}_{ij} - \mu_{ij}\right)^2}{a\left(\mu_{ij} - b\right)^2 + c} - \ln\left(a\left(\mu_{ij} - b\right)^2 + c\right) \right] \tag{13}$$

但し，$V_{ij}^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} \left(X_{ij}^{(k)} - \overline{X}_{ij}\right)^2$ は標本分散である．

ここから先は数値的に求める事になり，a, b, c を適当な初期値から出発して，徐々に改善させる．任意の a, b, c に対して，尤度を最大にする各 μ_{ij} は, 上式大括弧の中を微分して 0 とおけば求まり，具体的には以下の 3 次式に帰着する．

$$\begin{aligned} &-a^2\mu_{ij}^3 \\ &+a\left((1+3a)b - \overline{X}_{ij}\right)\mu_{ij}^2 \\ &+\left(-ab^2(1+3a) + a\left(\overline{X}_{ij}^2 + V_{ij}^2 - c\right) - c\right)\mu_{ij} \\ &+ab\left(ab^2 + b\overline{X}_{ij} - \overline{X}_{ij}^2 - V_{ij}^2 + c\right) + c\overline{X}_{ij} = 0 \end{aligned} \tag{14}$$

実数解が一つの場合はそれが求める μ_{ij} であり, 三つの場合は，元の式に数値的に代入する事で，どれが最大値を与えるかが判別できる．

3-2　変数変換

分散が平均の関数 f で表されるとすると，

$$\sigma_{ij}^2 = f\left(\mu_{ij}\right) \tag{15}$$

個々の観測データを Taylor 展開を用いた変換をする事で，分散を正規化することができる[2]．ここでは g を変換関数とし, 一時導関数 g' を持つと仮定して，$g(X_{ij})$を μ_{ij} の周りで Taylor 近似する．

$$g\left(X_{ij}\right) \approx g\left(\mu_{ij}\right) + g'\left(\mu_{ij}\right)\left(X_{ij} - \mu_{ij}\right) \tag{16}$$

$g(X_{ij})$の分散は 1 に正規化したいので，

$$\begin{aligned} \mathrm{Var}\left[g\left(X_{ij}\right)\right] &\approx \mathrm{Var}\left[g\left(\mu_{ij}\right) + g'\left(\mu_{ij}\right)\left(X_{ij} - \mu_{ij}\right)\right] \\ &= \left\{g'\left(\mu_{ij}\right)\right\}^2 f\left(\mu_{ij}\right) = 1 \end{aligned} \tag{17}$$

となり，$g'\left(\mu_{ij}\right) = \frac{1}{\sqrt{f\left(\mu_{ij}\right)}}$ の両辺を積分して，

$$g\left(X_{ij}\right) = \int_{C_0}^{X_{ij}} \frac{1}{\sqrt{f(t)}}\,dt + C_1 \tag{18}$$

を得る．C_0 及び C_1 は任意の定数である．f に上記の分散推定関数を代入し，C_0，C_1 を既存の対数変換と互換性があるように定める．

$$\begin{aligned} g\left(X_{ij}\right) &= \int_{1+b-c/4a}^{X_{ij}} \frac{1}{\sqrt{a(t-b)^2 + c}}\,dt \\ &= \frac{1}{\sqrt{a}} \ln \frac{1}{2}\left(\sqrt{\left(X_{ij} - b\right)^2 + c/a} + X_{ij} - b\right) \end{aligned} \tag{19}$$

これが求める変換式である．b=c=0 とすると，

$$g\left(X_{ij}\right) = \frac{1}{\sqrt{a}} \ln X_{ij} \quad \text{但し } X_{ij}>0 \tag{20}$$

となり，通常の対数変換になる．

4　検証

米国 National Center for Biotechnology Information (NCBI)のホームページの Gene Expression Omnibus (GEO) からダウンロード可能な公開データ GSE13 [3]を用いて，本手法の検証を行った（詳しくは[4]を参照）．このデータは Affymetrix GeneChip® Mu11K-A を用いて，6508 個の遺伝子について 5 条件で測定し，各条件下での繰返し測定数は各々6, 5, 5, 5, 5 である. 発現量は最小-6785, 最大 27413，平均 600.0，非正値の割合は 24.5%であった．

図 1(a)に，変換前の各遺伝子の各条件での測定データの標本平均と不偏分散のプロットを示す．大雑把には分散は平均の二乗に比例しているが，平均 0 付近でも分散は 0 にならず，平均が負のデータもある．実線は最尤推定から求めた分散推定関数で，データによく当てはまっている．図 1(b)は本手法で変換後のデータを同様にプロットしたもので，データを標本平均値に従って 100 グループに分割し，それぞれのグループ内で平均の平均，分散の平均を求めて繋いだものが実線，同様に平均の中央値，分散の中央値を求めて繋いだものが破線である．変換後の分散は平均に大きくは依存せず，ほぼ 1 に正規化されている．比較のため，通常の対数変換を行ったデータを図 1(c)に示す．非正値は対数変換できないため，それぞれの遺伝子についてどの条件であっても，一つでも非正値を含む遺伝子を除いてある（残遺伝子数=2365）．標本平均が小さくなるにつれ，不偏分散が増加する明らかな傾向が見られる．

図 2(a)に，等分散性の検定（Levene 検定）の結果を示す．これは，各遺伝子について，条件間で分散が有意に異なるかを検定したもので，任意の有意水準に対して，どれだけの遺伝子が等分散性を棄却されたかを表している．変換する前のデータは，変換後のデータに比べて棄却率が高い事が分かる．近似を使う検定のため，実際の有意水準を人工データから求め，併記してある．

図 2(b)に，正規性の検定（Shapiro-Wilk 検定）の結果を同様に示す．これは，各遺伝子の各条件について，データが正規性を満たしているかを検定したもので，任意の有意水準に対して正規性を棄却された率を表している．変換前後で正規性は殆ど変わらず，良好な事が分かる．これも近似を使う検定のため，実際の有意水準を人工データから求めて併記した．また，対数正規分布に従う人工データの棄却率も参考に併記した．

5　まとめ

我々は，事前のパラメータ設定が一切必要なく，非正値も扱え，等分散性を改善する，対数変換に似た変換方法を開発し，公開データで検証した．また，正規性は変換前も良好で，変換後も悪化しなかった．この変換は従来の対数変換に代わるもので，多くの解析の前処理として有用であると思われる．

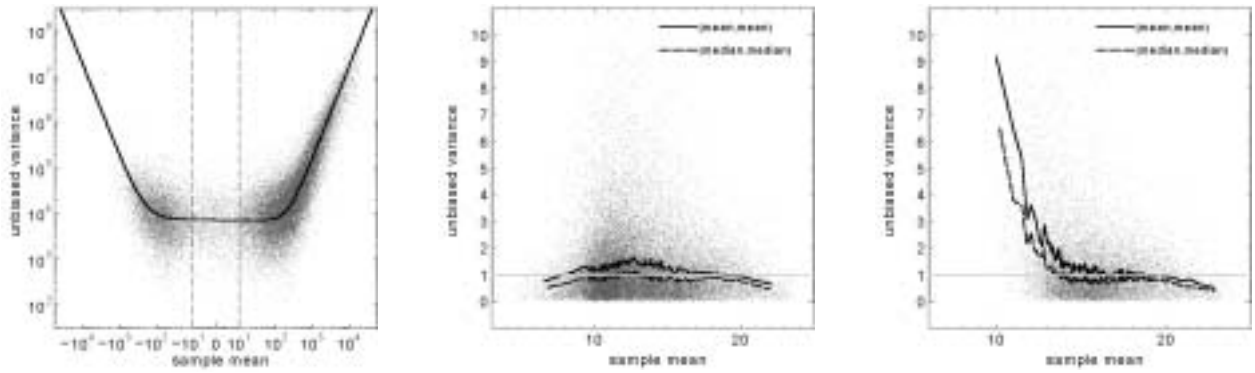
参考文献

[1] Affymetrix, “*Microarray Suite User Guide*,” Santa Clara, 2000.

[2] Rao, C.R., “*Linear Statistical Inference and Its Applications*, (2nd ed.),” John Wiley & Sons, New York, 1973.

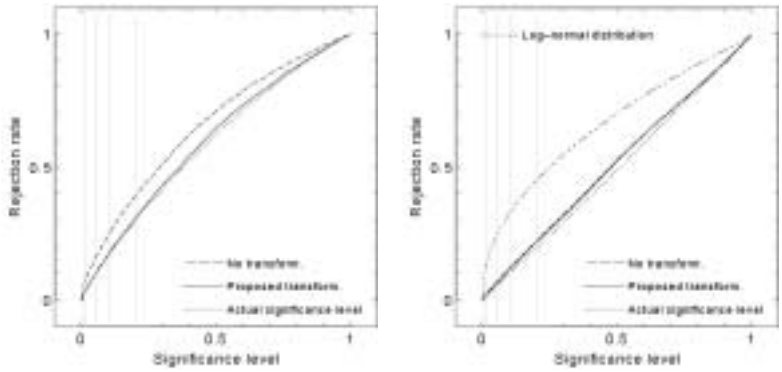
[3] Hoffmann, R., et. al., “Changes in gene expression profiles in developing B cells of murine bone marrow,” *Genome Res.*, **12**(1), 98-111, 2002.

[4] Inoue, M., et. al., “Preprocessing of DNA microarray data: a mathematically suitable alternative to the conventional logarithmic transformation,” *Bioinformatics*, (submitted).



(a)　(b)　(c)

図 1　(a)変換前，(b)提案手法で変換後，(c)対数変換後，の各遺伝子の各条件における発現量の標本平均と不偏分散のプロット．(a)の実線は分散推定関数 $f(\mu)=0.1768(\mu- 27.47)^2+6955$．



(a)　(b)

図 2　(a)等分散性，(b)正規性の検定結果．変換前を破線で，変換後を実線で，実際の有意水準を点線で示した．(b)では，変換前と変換後の線は殆ど重なっている．2 点破線は対数正規分布に従う人工データでの棄却率を示す．