

Title	非計量多次元尺度構成法とその生体生命情報解析への応用(非線形科学と統計科学の対話,研究会報告)
Author(s)	田口, 善弘
Citation	物性研究 (2008), 91(2): 137-143
Issue Date	2008-11-20
URL	<a href="http://hdl.handle.net/2433/142694">http://hdl.handle.net/2433/142694</a>
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

# 非計量多次元尺度構成法とその生体生命情報解析への応用

中央大学 理工学部/理工学研究所 田口善弘<sup>1</sup>

非計量多次元尺度構成法は、もともと社会科学で使われていた方法であるが、我々はこの10年ほどの間、生体生命情報（いわゆるバイオインフォマティクス）への応用を試みてきた。今回はその中から、分裂酵母の細胞分裂周期のマイクロアレイ実験の解析結果を報告する。

## 1 非計量多次元尺度構成法とは

非計量多次元尺度構成法 [1](nMDS) は、一般には自由度の低減を行うための多変量解析手法であるとみなされている。非計量多次元尺度構成法が対象とするのは  $N$  個の事物間の非類似度  $\delta_{ij}$  の組（ここで  $i, j = 1, \dots, N$  は事物に付けられた任意の番号である）間の大小関係、

$$\delta_{ij} > \delta_{kl} \quad (1)$$

である ( $k, l = 1, \dots, N$  もまた  $N$  個の事物の番号である)。多次元尺度構成法の目的は

$$d_{ij} > d_{kl} \quad (2)$$

を満たすような（つまり、大小関係が一致するような）、計量空間における  $N$  個の事物の配置（付置）を得ることにある（ここで  $d_{ij}$  は所与の計量空間における  $ij$  間の距離である）。非計量という名前は順序のみを情報として取り扱い、具体的な「量」を対象としない、ということから来ている。わざわざこの様な回りくどいやりかたをとる理由は、 $\delta_{ij}$  が非類似度を定量的に与えていることが疑わしい場合が多いからである。

実際には式 (1) と (2) を同時に満たすことは不可能である。そこで「なるべく」式 (2) を満たすような付置を得よ、ということになる。この「なるべく」という基準がなんであるのかについての意見の統一がないために、様々な流儀の nMDS が存在している。我々は、屋上屋を重ねる様ではあるが、「なるべく」の基準として「順位の差の2乗和が最小（つまり、 $d_{ij}$  と  $\delta_{ij}$  の間の Spearman の順位相関係数が最大）」という基準を導入して新たな nMDS のアルゴリズムを提案した [2]。この方法は、従来の方法より速く、また、極小解に捕まりにくいという利点があり、大きな  $N$  に対して nMDS を用いる場合の強力なアルゴリズムとなっている。手法の詳細は文献 [2] にゆずる<sup>2</sup>。

<sup>1</sup>E-mail: tag@granular.com

<sup>2</sup>これら文献はいずれもインターネット上で無償/有償にて download 可能である

## 2 生体生命情報学

本研究の手法は生体生命情報学（バイオインフォマティクス）の手法を踏襲していることになると思われる。ただ、個人的には、道具としてはその様な手法をとっていても目指すところは微妙に異なっているように感じているが今回はその詳細は割愛する。

### 2.1 データ

今回、用いるデータは、マイクロアレイという名称で呼ばれている実験手法で得られたデータである。マイクロアレイは、簡単にいうと、ゲノムからの転写物の相対的な転写量を定量的に測定する手法である。具体的には、個々の転写物の塩基配列と相補的な塩基配列断片がプレート上に配置されているものがあり、それに細胞抽出液を「注ぐ」ことにより、断片に転写物を結合させることで、転写物の量を測定する。従って、まず、どのような転写物の量を測定するかを決め、それらを網羅した相補鎖の断片を用意し、プレート上に配置する必要がある。本研究で解析するデータを得るために用いられたプレートはいわゆる遺伝子（=蛋白質に翻訳されることが解っている転写物）の転写量を測定するためのプレートであり、対象生物としては分裂酵母（fission yeast）であり、5000弱の遺伝子の転写量を測定することができる。この様なプレートを用いて、複数のグループ [3, 4, 5] が独立に細胞分裂周期のマイクロアレイ実験を行った。具体的には酵母の集団をある特定の細胞周期に同期化させた後、フリーランを行わせ、数周期に渡ってある時間間隔でマイクロアレイによる転写量の相対変化を計測する。目的は、細胞分裂の周期の間に、どの遺伝子がどれくらい発現しているか、を調べることにある。本研究で対象としたデータは Elutriation（遠心分離器によるサイズスクリーニング）を用いて細胞分裂周期を同期化したデータである。分裂酵母は細胞のサイズが2倍になると2個の細胞に分裂するという単純な分裂機構を持っているので、同じサイズの酵母は細胞分裂周期の同じ相にいたることが期待される。

### 2.2 従来の研究

分裂酵母の分裂周期（細胞分裂に必要とする時間）は実験条件によって変化するが、分裂の様子を観測していれば容易に知ることができる。この周期  $T$  を用いて、転写量  $x(t)$  の時間  $t$  による変化が

$$x(t) \simeq \cos(\omega t + \delta) \quad (3)$$

(但し、 $\omega = 2\pi/T$ )、に近い遺伝子を細胞分裂周期遺伝子とみなして細胞分裂に関係する遺伝子を決めよう、というのが従来の方法である（いわゆる、正弦回帰）。このやりかたは一般に広く受け入れられていて、観測グループ [3, 4, 5] もこの方法でデータを解析し、数百個の遺伝子を見付けている。更に、その後、このラインに沿った再解析もなされている [6]。しかし、常識的に考えて、転写量が時間の関数として正弦波になっている理由もないし、実際、実験グループが選んでいる

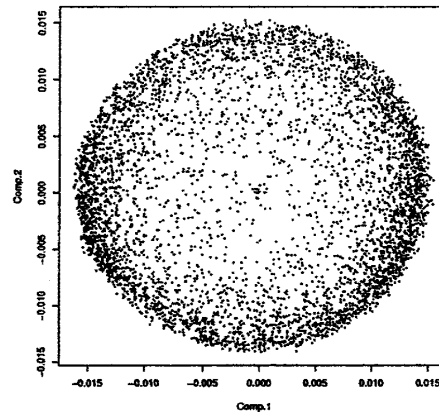


図 1: nMDS により 2 次元空間に埋め込まれた遺伝子発現プロファイル。個々の点がプロファイルに相当している

遺伝子のセットは一致がよくない。なにより、数百個であるという根拠が明解でない。なぜ、数十個でも千個でもないのかについての説明がないのである。

### 3 nMDS の応用

この問題に我々は nMDS を用いて挑戦してきた [7]。ここではその最新の知見について簡単に報告したい。

#### 3.1 方法

nMDS を用いて、各遺伝子の発現プロファイル（転写量の時系列データ）の関係を可視化する。ここで nMDS で関係性が可視化される「事物」としては、個々の遺伝子（の発現プロファイル）ということになる。nMDS を用いるには時系列データ間の「非類似度」を定義しなくてはならないが、ここでは我々は時系列データ間の相関係数の符号を反転させたものを用いることにした（相関係数の値が大きい程、時系列は似ていることになる。非類似度にするには大小関係を逆転させねばならないので、逆符号をとった）。埋め込み空間としてはまず、2次元を採用した。

このような処理を行うと一般にリングが出現することがよく知られている（図 1）。これは時系列データ間に、なんらかの 1 次元的な周期的な関係性があることを暗示している。これが細胞分裂周期であればいうことはない。式 (3) の様な生物学的な根拠が皆無である式を使わずに、細胞分裂周期に沿って遺伝子=発現プロファイルを並べることができたことになるからである。

## 3.2 結果

### 3.2.1 再現性

まず、図1のリングに沿った極角度の再現性について議論しよう。リングに沿った遺伝子の並び順（角度）に再現性がなければ、そもそも、解析が無意味である。ここでは参考文献 [3, 4] からとったそれぞれ独立な2つの Elutriation 実験を nMDS で解析した場合の遺伝子の角度依存性の再現について議論する。図2より明らかな様に、リングに沿った「角度」には再現性があり、それは

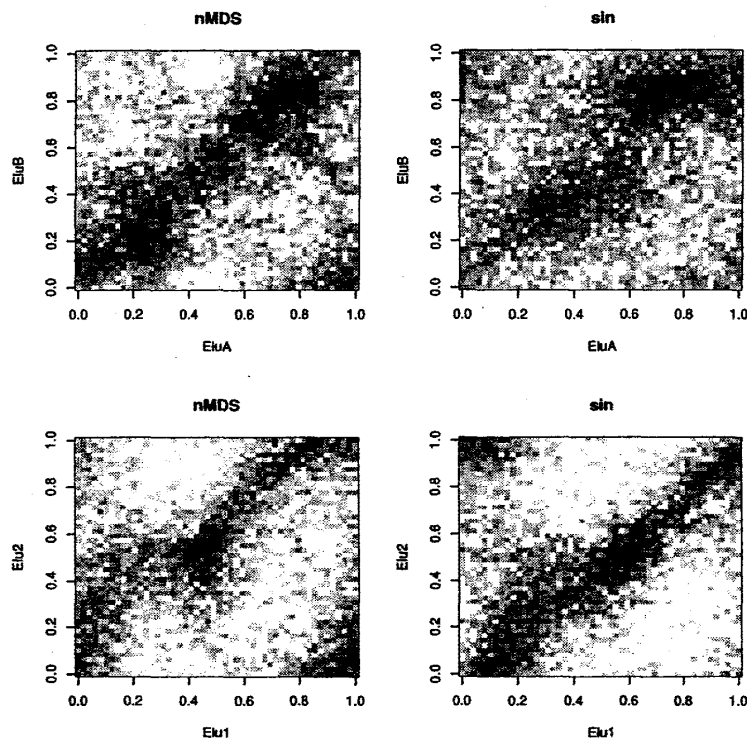


図2: 左: 図1に見られるような、リングに沿った遺伝子の角度依存性の再現性。右: 式(3)を仮定した場合の位相 $\delta$ の再現性。EluA, EluBは参考文献[3]より、Elu1, Elu2は参考文献[4]より

式(3)で計算された位相と同程度の再現性を持っていることが解る。参考のため nMDS で求めた極角度と正弦回帰(3)式で計算した位相との整合性もプロットしてみた(図3。特に参考文献[4]の場合は非常によく一致している。つまり、これらの場合は式(3)を仮定した解析は悪くない、と言える。この様な図をみせるとそもそも、正弦回帰でいいのではないか、nMDSを使う意味がないのではないか、というコメントが来そうであるが話はそう単純ではない。これらの図には「全ての遺伝子の位相」が描かれていることに注意すべきである。つまり、個々の遺伝子の発現プロファイルが正弦関数であろうがなかろうが強引に正弦回帰した場合の位相に「意味」があるということである。実際、遺伝子発現プロファイルの多くは通常の検定を行った場合、式(3)である可能性が有意でない時系列である。にもかかわらず求めた位相には意味がある。実際、詳しく見ると式(3)の有意性は棄却されてしまうのに、位相 $\delta$ は十分な精度(1周期に換算した場合の誤

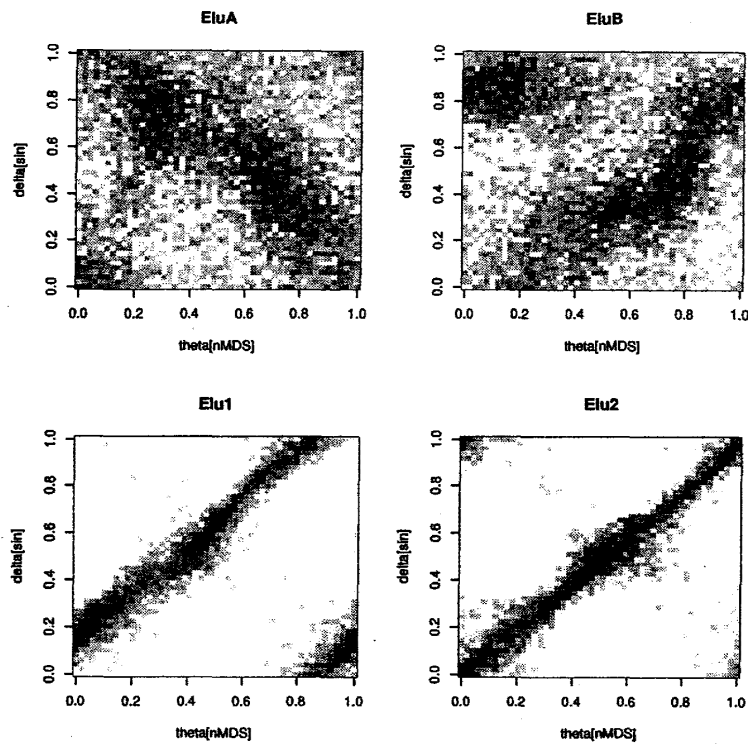


図 3: nMDS で求めた位相 (極角度: nMDS) と式 (3) を仮定して求めた位相の比較。

差が5%程度) という一見矛盾したことが頻繁に起きていることが解る。このことは通常の手続きによるデータ解析に重大な反省を迫るものであると思われる。つまり、式に意味がないのにパラメータには意味がある、という場合をどう扱うかという問題である。式の形を仮定しない nMDS の場合にはこの問題が顕在化しないのでうまく行っているように見えると思われる。

### 3.2.2 細胞周期遺伝子とは?

正弦回帰 (式 (3)) によく当てはまっているものが細胞周期遺伝子だという考え方に問題があることを全小節でみた。それではどうするか? ここでは整合性という観点から遺伝子を選択してみた。マイクロアレイなどの網羅的な研究以前から、数十個の遺伝子は細胞分裂に関与していることが既に知られていた。そこで、これらの遺伝子との位相差 (あるいは、これらの遺伝子から見たときの相対的な周期の差) が図 2,3 でとりあげた4つの実験においてなるべく不変である遺伝子を選ぶ、という観点から遺伝子を選んでみた。つまり、既知の細胞周期遺伝子と協同的に発現する遺伝子を探したわけだ。この様にして選んだ遺伝子の位相を4つの実験について比較した図 4 を載せておく。注意すべきことは実験間での整合性を直接要請したのではないにも関わらず、位相の実験依存性が非常にすくなくなっていることである。比較のために載せた Peng ら [5] が正弦回帰 (3) を仮定して求めた位相もよく一致している。このことから、やはり、位相と言うものは正弦回帰を仮定して求める、というより、遺伝子発現プロファイルの「差」の大小から決まって

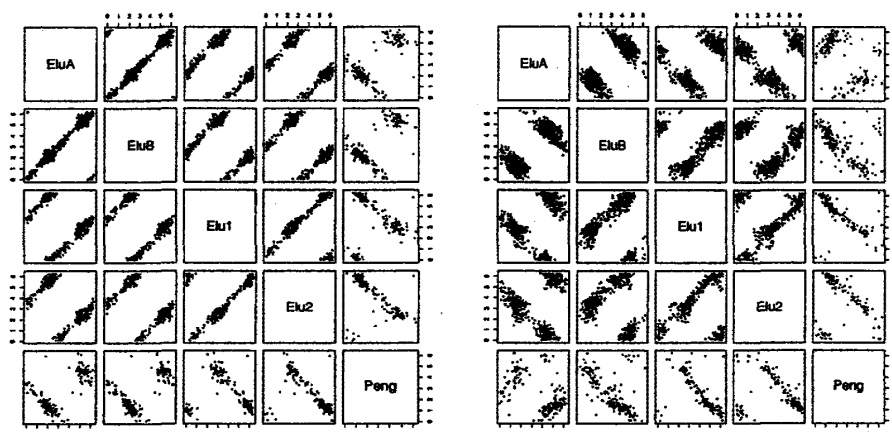


図 4: 「細胞周期遺伝子」の位相の実験間整合性。左: 式 (3) を仮定した場合。右: nMDS を使ってリングに沿った極角度を使った場合。Peng *et al*[5] が式 (3) を仮定して求めた位相を比較のために載せた。

いる (つまり、形が似ている程、「位相」が近い) らしいことが解る。この観点からは nMDS の様な手法がより望ましい。

#### 4 結論

本稿では nMDS を生体生命情報データの解析に応用することで、従来の統計的な解析では矛盾を孕むような場合でも問題なく扱うことができることを示した。nMDS は生体生命情報データの解析に有用な方法であることが期待される。更にいろいろなデータへの応用を試みたい。

#### 参考文献

- [1] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory And Applications*, (Springer, 2005).
- [2] 田口善弘, 大野克嗣, 横山和成「非計量多次元尺度構成法への期待と新しい視点」統計数理、49 (2001) 133; Y-h. Taguchi and Y. Oono, *Nonmetric Multidimensional Scaling as a Data Mining Tool —New Algorithm and New Targets—*, in *Geometrical Structures of Phase Space in Multi-Dimensional Chaos: Applications to Chemical Reaction Dynamics in Complex Systems*, (Advances in Chemical Physics, vol 130, Part B), eds. M. Toda, T. Komatsuzaki, T. Konishi, R. S. Berry, S. A. Rice (Wiley Interscience, 2005), 315-351; Y-h. Taguchi and Y. Oono, *Relational patterns of gene expression via nonmetric multidimensional scaling analysis*, *Bioinformatics*, 21 (2005), 730-740.

- [3] Anna Oliva, Adam Rosebrock, Francisco Ferrezuelo, Saumyadipta Pyne, Haiying Chen, Bruce Futcher, and Janet Leathwood, The Cell Cycle-Regulated Genes of *Schizosaccharomyces pombe*, *PLoS Biology* **3** (2005) e225.
- [4] Gabriella Rustici, Juan Mata, Katja Kivinen, Pietro Lió, Christopher J Penkett, Gavin Burns, Jacqueline Hayles, Alvis Brazma, Paul Nurse, Jürg Bähler, Periodic gene expression program of the fission yeast cell cycle, *Nature Genetics* **36** (2004) 809-817.
- [5] X. Peng, K. M. Karuturi, Miller Lance D, Lin K, Jia Yonghui, Kondu P, L. Wang, L. Wong, Liu Edison Tak-Bun, M. Balasubramanian, Liu Jian-Hua, Identification of cell cycle-regulated genes in fission yeast, *Mol. Biol. Cell*, **16** (2005) 1026-1042 .
- [6] Marguerat S, Jensen TS, de Lichtenberg U, Wilhelm BT, Jensen LJ, Bahler J., The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast* **23** (2006) 261-277.
- [7] Y-h. Taguchi, Gene arrangement for cell division cycle microarray experiments without sinusoidal fittings, IPSJ SIG Technical Report, 2006-BIO-6, (2006) 173-180; Y-h. Taguchi, Detecting Cell Cycle Regulated genes in *S. pombe* with Non-metric Multidimensional Scaling without sinusoidal fittings IPSJ SIG Technical Report, 2006-BIO-3 (2005) 59-66.