非定常時系列の教師なしクラスタリング：位相的データ解析による分析

黒木 裕鷹

2018 年 4 月 27 日

1　はじめに

膨大な特徴量をもつ高次元データ解析において，どのように高次元の特徴量を低次元で表現するのかは重要な研究課題である．近年，位相的データ解析による高次元データ解析が盛んに提案されており，複雑な高次元データの位相的情報を可視化する手法として注目を集めている．

位相的データ解析では、主としてパーシステントホモロジー [1] という、ホモロジーの誕生と消滅に着目することにより、データの位相的情報を抽出する手法が用いられる。このパーシステントホモロジーを可視化したパーシステントダイアグラム [2] を解析することによる知識発見や、変数間のネットワーク構造の解明に役立つことが期待されている。

位相的データ解析を利用した研究としては、複数方向からのシルエット画像からパーシステントホモロジーを構成し歩行識別を行った [3]、[4] による病気に寄与する遺伝子の解析、[5] による金融危機の早期察知，[6] による金融資産間の相関構造の推移に関する分析などがあり、その応用分野は多岐にわたる。

現在，世界の様々な工場では産業用ロボットが用いられている．産業用ロボットの導入の目的は、人間には負担の大きい作業の代替のためであり、例えば、重量物の運搬や作業場に霧散している粉塵の吸入にリスク回避等を目的としている。産業用ロボットはこのような安全性の問題を解決するだけでなく，単純作業の代替などによる人件費削減など、経済性や効率性においてもメリットがある．一方、ロボットには故障のリスクがあり，ただひとつのロボットの故障が生産ライン全体に影響を及ぼし得る．故障を未然に回避するためにはメンテナンスなどの保守作業を行うことが重要であるが，メンテナンスや部品交換にも費用や時間などコストがかかるため，適切なタイミングで行われることが望ましい．

自動車メーカーの M では，生産ラインで使用するアームロボットの故障を未然に防ぐため，その減速機の交換を経験に基づくタイミングで行っている．交換のタイミングが遅すぎれば故障を招くが，早すぎても交換に時間がかかる分パフォーマンスが低下することとなる．本調査では，ロボットアームをモニタリングした振動のセンサーデータを用い，減速機交換前後のデータに明確な違いがあるかどうかを調査した．ロボットの故障により生産ラインを止めることはあってはならないため，故障直前のデータは得ることが出来ないことに留意しなければならず，教師無し学習によるクラスタリングを行うことを目的とした．

M より提供されたデータは一つのロボットアームにつき，減速機交換前後それぞれで 5 秒間の計測を 10 回行った 1 次元の振動データである．ロボットアームはそれぞれ挙動が異なるため，その主要な振動はアームごとに異

なっている．また，各アームの行動 1 セットは 5 秒間ではないため，各計測ごとにより行動セットの中の計測する部分が異なっている．

複数のロボットアームの振動は高次元時系列として与えられる。また、各アームやセンサー取り付け位置に応じて、振動パターンの形状は様々な周期的変動パターンを持つ。ことのうな非定常高次元高頻度時系列のクラスタリングに対する統一的な統計分析手法は存在しない。多変量解析における階層的クラスタ分析では、クラスタリングされる対象間の類似度もしくは非類似度が必要である．時系列を対象としたクラスタリングにおいて，最も広く使用されている非類似度はユークリッド距離，Dynamic Time Warping(DTW) [7]，CORT [8] などである．スペクトル密度関数を特徴とするクラスタリングは、などによって提案されている。DTW は 2 つの時系列データの異なる 2 時点間全ての誤差を計算し，一定の制約の元でそれらの合計が最小になるような一致を求めるアルゴリズムである．DWT や CORT による分析は、例えば、、や。。の分析を対象に行われ、、、などの有用性が示されている。DWT を用いることのメリットは、DTW は二つの時系列の周期や長さが異なる場合でも非類似度を算出することができることが挙げられる．DTW により挙動の計測開始時点が異なる問題は多少解決されるが，振動の形状の大部分はロボットアームの挙動によって異なるため，減速機交換前後の差異はクラスタリングの結果に表れない．また，未知音源分離などで用いられる FastICA [9] などの独立成分分析 (ICA) のアルゴリズムは、主要な振動と微細な振動を分けることを可能にするかもしれない。しかし，同時点の複数観測を必要とする他、未知音源の数を指定する必要があり、本データには適用することはできない．時系列の同時性を利用することなく，かつアームごとに異なる主要な振動に影響されない特徴量によって教師無し学習を行うことが課題である．本調査では，以上のような課題を解決するため，データの位相的な特徴に注目した．データの位相的特長を分析することにより、故障につながり得るロボットアームの劣化を抽出することを目的とする。

近年着目されている Topological Data Analysis(TDA) は複雑なデータを扱う上で強力であり，パーシステント・ホモロジー [1] とその表現であるパーシステントダイアグラム [2] をはじめとするその手法はデータの位相的特徴を抽出し，新しい知見を与える．また TDA の分野で広く行われているように，時間遅れ座標を用いて時系列を多次元に埋め込むことで位相的特徴を抽出する足掛かりとした．時間遅れ座標による埋め込みはダイナミカルシステムの分野で，状態空間の復元を目的に広く用いられている．パーシステント・ダイアグラムを比較するために Wasserstein 距離 [10] がしばしば用いられるが，本調査で扱うような高頻度データでは計算量の観点で現実的ではない．また，[11] の提案する Betti sequence はデータの位相情報を 1 次元の系列として要約する．そこで本調査では，観測の上述した観測の長さや観測開始時点が異なる問題を解決するためにセンサーデータの位相情報を要約した Betti sequence をもとめ，DTW に基づく全結合型階層的クラスタリングを行うことにより，減速機交換前後の振動を分析した．

本レポートの構成は次のようである．まず 2 節では本調査で使用した TDA の手法やクラスタリング手法について述べ，3 節では実際のデータ解析とその結果を示す．4 節では考察を行うと共に今後の展望について触れる．

2 Topological Data Analysis(TDA) と時系列クラスタリング

位相空間とは、集合に距離関数を定義することなく、直接開集合となる部分集合を定めてできる空間である。位相的データ解析では、距離関数に縛られないやわらかい幾何情報を扱い、データの情報を抽出する。その主要な手法は mapper [12] とパーシステント・ホモロジー [1] である．これらはノイズを含む複雑なデータセットから何か新たな知見を得る目的でしばしば用いられてきた．本調査ではパーシステント・ホモロジーを利用する．

2.1 パーシステント・ホモロジー

ユークリッド空間の有限点集合を X，点 $x_i \in X$ を中心とした半径 r の球の和集合を $B(X;r) := \bigcup_{i=1}^{n} B(x_i;r)$ とする．ただし，$B(x;r) = \{y \in M | d(x,y) \leq r\}$（$d(x,y)$ は x, y のユークリッド距離）とする．球の和集合を半径パラメータ r で集めた集合 $\mathbb{B}(X) := \{B(X;r)\}_{r \geq 0}$ をここでは X のフィルトレーションという．$r \leq a$ ならば包含関係 $B(X;r) \subset B(X;a)$ があるため，ホモロジー群間の射 $u_r^a : H_q(B(X;r)) \to H_q(B(X;a))$ を包含写像から誘導する．このとき，ホモロジー群の系列

$$H_q(\mathbb{B}(X)) : \cdots \to H_q(B(X;r)) \overset{u_q^b}{\to} H_q(B(X;a)) \to \ldots (r \leq a)$$

を X の q 次元パーシステントホモロジーという．また、パーシステントホモロジーは分解定理 [13] により、適切な区間表現 $\mathbb{I}[b_i, d_i]$ を通じて

$$H_q(\mathbb{B}(X)) \cong \oplus I[b_j, d_j](b_j \leq d_j)$$

で一意に表現される．ここで、区間 $I[b_j, d_j]$ は q 次ホモロジー $j(j = 1, \ldots, J_q)$ が発生 (b_j) してから消滅 (d_j) するまでの r の区間である。これをユークリッド空間 $\mathbf{R}^2$ 内にプロットした散布図

$$D_q(X) := \{(b_j, d_j) | j \in J_q\}$$

を X の q 次元パーシステント・ダイアグラムという．パーシステント・ダイアグラムの元 (b_j, d_j) はホモロジーの生成元の発生時間 (Birth time) を b_j，消滅時間 (Death time) を d_j と記録しているものと解釈できる．

q 次のパーシステント・ダイアグラムは 2 次元の散布図として表され，その解釈や扱いが困難である．そのため，様々なダイアグラムの要約が提案されてきた．最も単純な要約が最大パーシステンスであり，$\max_j(b_j - d_j)$ で表され、最も特徴的なホモロジーの持続性を表意味する。

2.2 パーシステント・バーコード

前節で定義したダイアグラム $D_q(X)$ のホモロジー j に対して，次の $s_j(r)$ を定義する．

$$s_j(r) = \begin{cases} 1 & (b_j \leq r \leq d_j) \\ 0 & otherwise \end{cases}$$

$s_j(r)$ を並べてプロットしたものをパーシステント・バーコードという．

また，[11] はバーコードを単純に扱い，機械学習の枠組みで扱いやすくするために次で提案される Betti sequence, $BS(X)$ を提案している．

$$BS(X) = \sum_j s_j(r)$$

Betti sequence は半径 r のときにいくつのホモロジーが存在しているかを表す 1 次元の系列となる．

2.3 1 次元時系列データに対する TDA

時系列から複数次元の有限点集合を構成する方法として，遅れ時間座標を利用した埋め込み (embedding) がある．埋め込みはアトラクタを再構成するために非線形ダイナミカルシステムの分野で盛んに利用されている．長さ N の時

系列 $x_1, x_2, \ldots, x_T$ から適当な遅れ時間 τ ごとの d 個の測定値を取り出し $V(t) = (x(t), x(t+\tau), \ldots, x(t+d-1))$, $(t = 1, 2, \ldots, T)$ の d 次元有限点集合を得る．この有限点集合が元の k 次元力学系の埋め込みになるための十分条件として，$d \geq 2k+1$（ターケンスの埋め込み定理 [14]）が知られている．

また，有限点集合の次元 d が大きくなったときに $\mathbb{R}^d$ の n 個の点集合のドロネー三角形分割の計算量は $\mathcal{O}(n^{\frac{[d]}{2}})$ になりうる [15], [16]．そこで，位相情報を保持しながら計算を簡略化するため，十分多次元に埋め込んだ後主成分分析により 3 次元に次元削減するアプローチが提案されている [17].

2.4 時系列クラスタリング

時系列間の非類似度を算出する際，用いられているのが [7] で提案された Dynamic Time Warping(DTW) である．DTW は三角不等式を満たさないため距離ではないが，比較的少ない計算量で要素数の異なる系列同士の距離のような量を求める手法である．

また本調査では DTW に基づき，完全連結方による階層型クラスタリングを行った．完全連結法とは，クラスタ間で最も類似度が低いデータ間の距離をクラスタ間の距離にする方法である．

3 データ解析

3.1 データの概観

本調査では，自動車メーカー M の産業用ロボットアームの振動データを扱う．15 種類のロボットアームに対し，減速機を交換する前と交換した後をそれぞれ 10 回モニタリングした合計 300 系列のデータである．計測の基本単位は 100 ミリ秒であり，計測の長さは 5 秒間である．ロボットアーム 15 種類の減速機交換前後のデータのそれぞれ 10 回分の観測を連結し，以下の図 1 にプロットした．このように，いずれのロボットアームにおいてもその時系列は明らか非線形構造を持つと同時に交換前後の系列の挙動は類似しておりは，アームごとに特有の周期的な振動を持つことがわかる．減速機交換前後の特徴を目視で判断することはできない．さらに交換前後で計測開始時間が異なるものもあり，単純な比較が難しいことがわかる．

3.2 データの位相的特徴

ロボットアームによる振動のスケールを調整するため，以下の分析は予め全系列を標準化した上で行った．1 次元の振動時系列から位相情報を取り出すために，遅れ時間座標への埋め込みを行った．その際，遅れ時間単位は $\tau = 1$ とした．またロボットアームの構造や状態変数の数が明らかでないことから，埋め込み次元にはロボットアームが 3 次元であることを考慮し，ターケンスの埋め込み定理 [14] を参考に $d = 3 \times 2 + 1 = 7$ とした．さらに [17] で提案されているように主成分分析を用い第 3 主成分までを取り出し 3 次元の有限点集合を抽出した．アルファ複体のフィルトレーションを用いて抽出した点集合の $0, 1, 2$ 次元パーシステントホモロジーを計算した．0 次のホモロジーは連結を、1 次は穴を、2 次は空洞であるので、それぞれのパーシステントホモロジーはその発生と消失を表す。またその要約として各次元の Betti sequence を計算し，以下の図 2, 3 に示した

Betti sequence は、その時（半径）にいくつのホモロジーが存在しているのかを表すのであった。図 2, 3 より，ロボットアームごとにホモロジーの数の推移が異なり、位相的特徴が抽出されている。また，2 次の Betti sequence

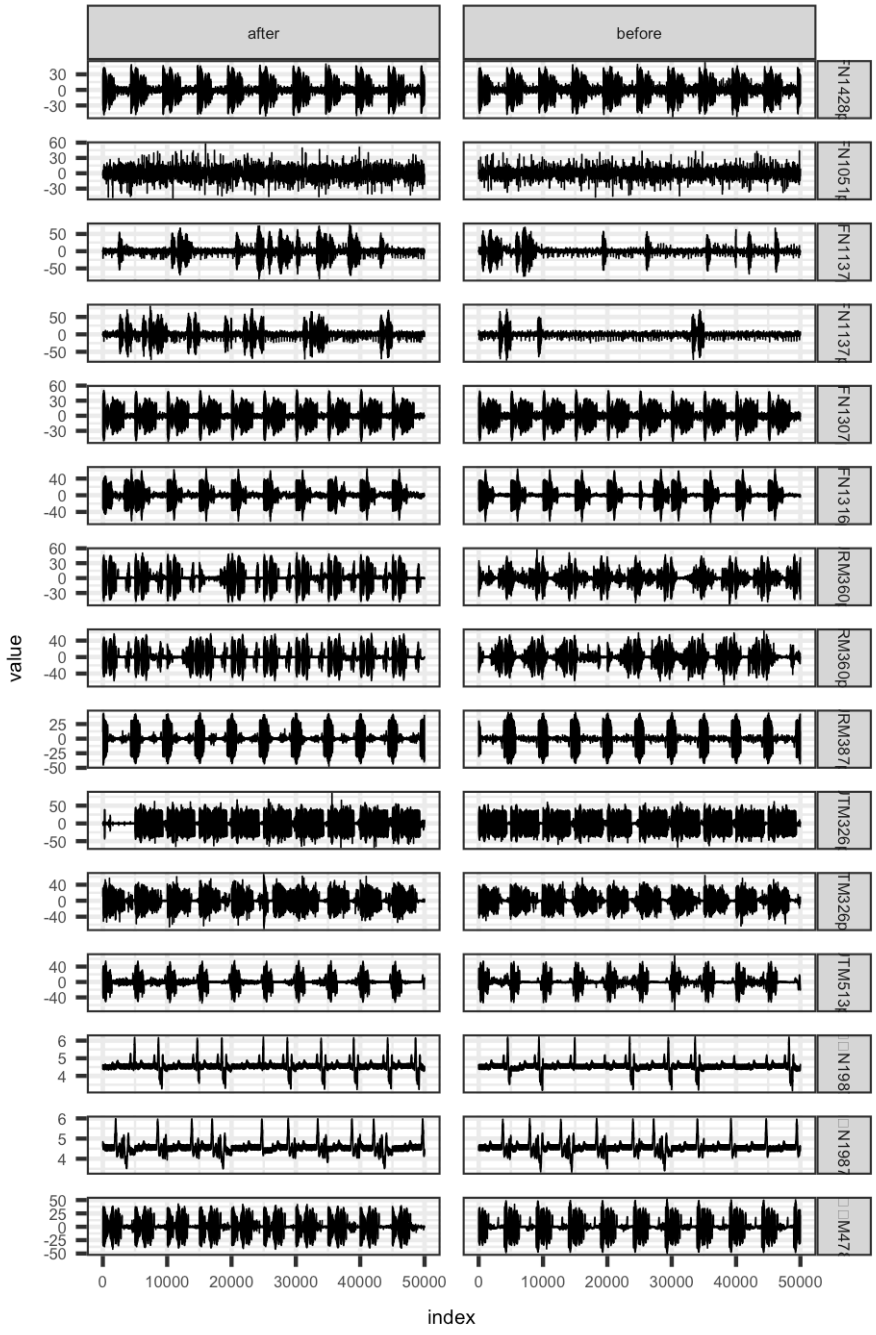

図 1: センサーデータの時系列プロット

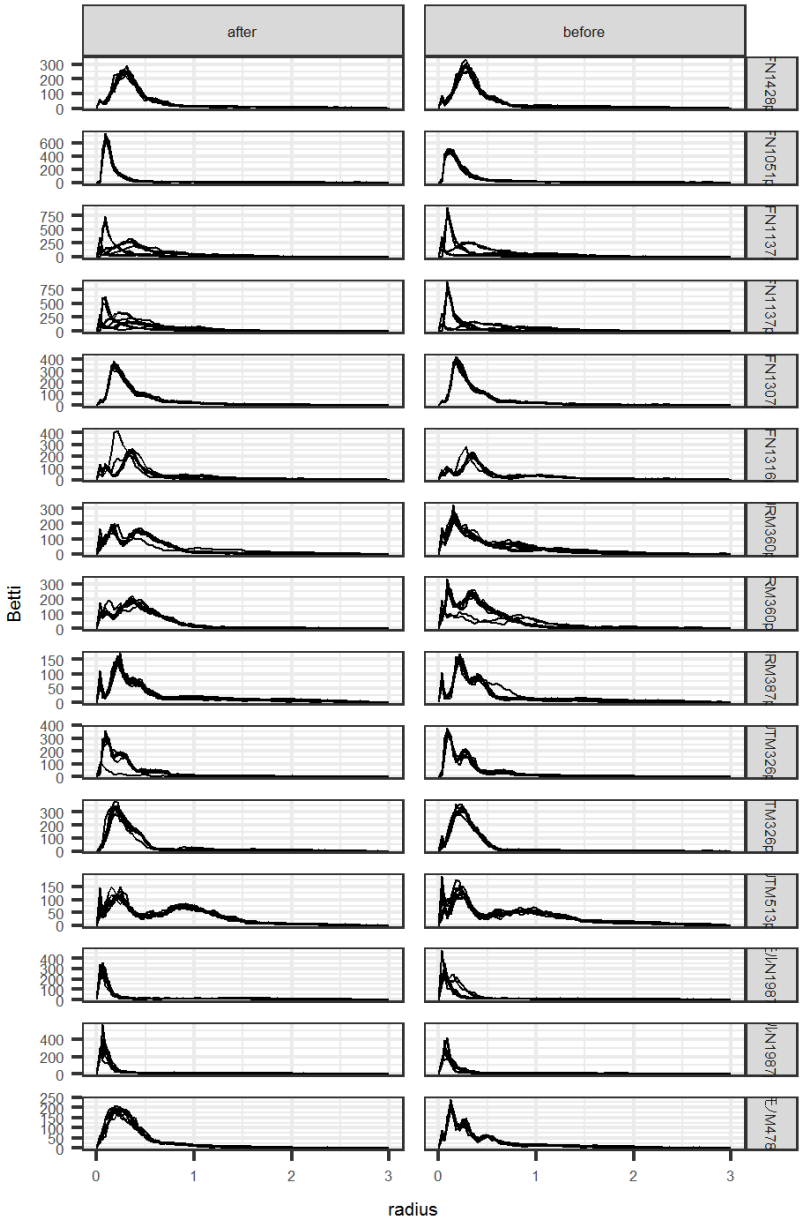

図 2: 1 次の Betti sequence

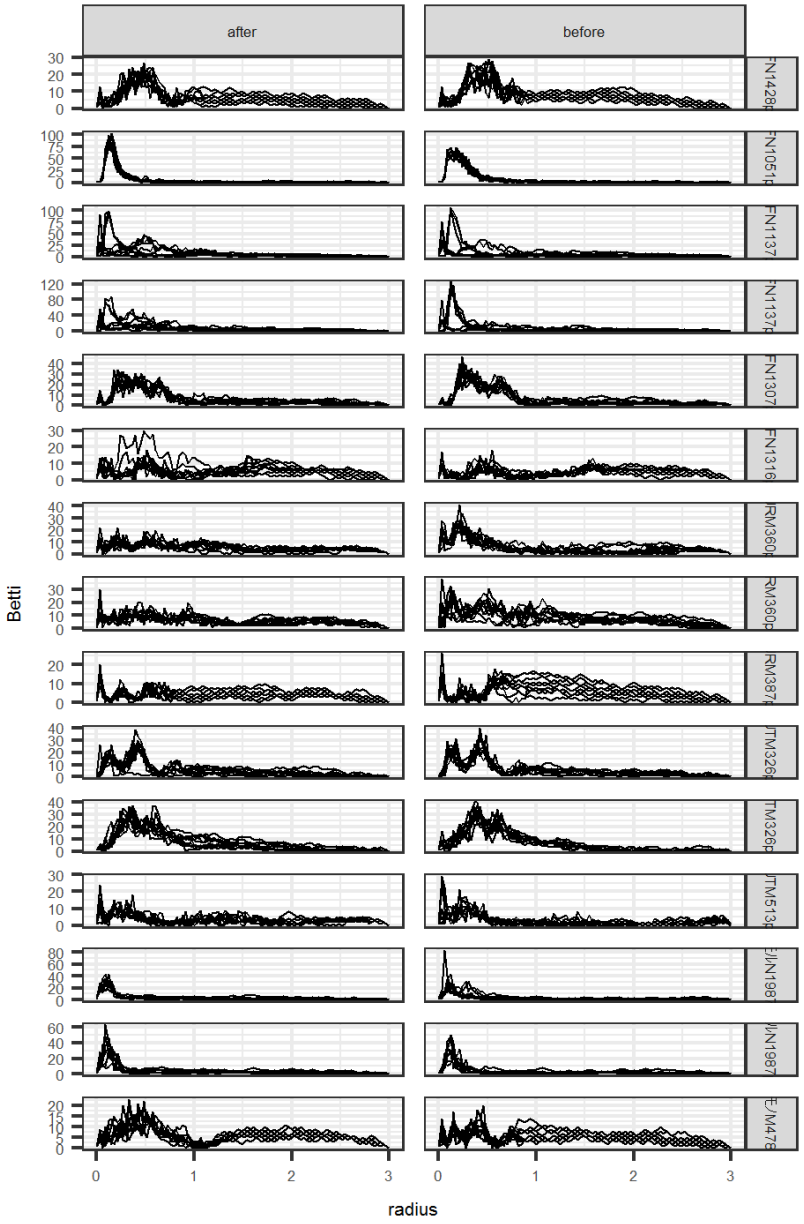

図 3: 2 次の Betti sequence

は 1 次のものに比べて観測ごとのばらつきが大きくなっていることが見て取れる．観測ごとのノイズに影響を受けている可能性があり、2 次の Betti sequence については扱いに注意が必要である。しかし、ロボットアームごとにその概形が異なることから、位相的特徴は抽出されている。

3.3 DWT に基づいた時系列クラスタリング

ロボットアームごとに減速機交換前 10 系列，交換後 10 系列の計 20 系列をクラスタリングする．観測ごとのばらつきが少ない 1 次のベッチ数をもとに DTW を比類似度とした階層的クラスタリングを行った．クラスター間の距離は完全連結法を用いて算出した．そのデンドログラムを図 4 から図 18 に示した．

図 4 から図 18 より，いくつかのロボットアームでは明確に減速機交換前後のクラスタが構成できたことが分かる．またそうでないアームでも細かい部分では減速機交換前後のクラスタができていることが分かる．

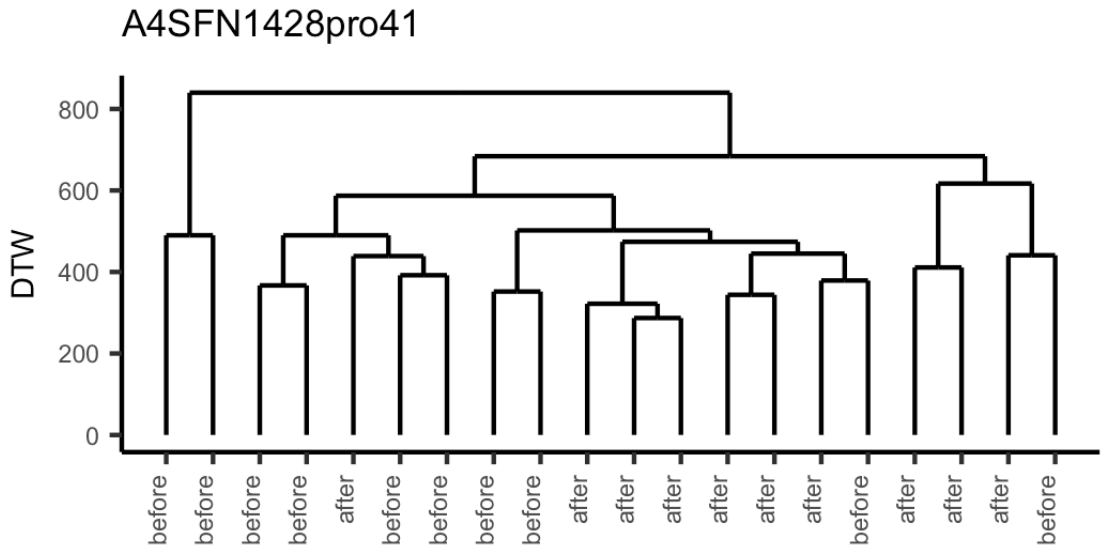

図 4: デンドログラム 1

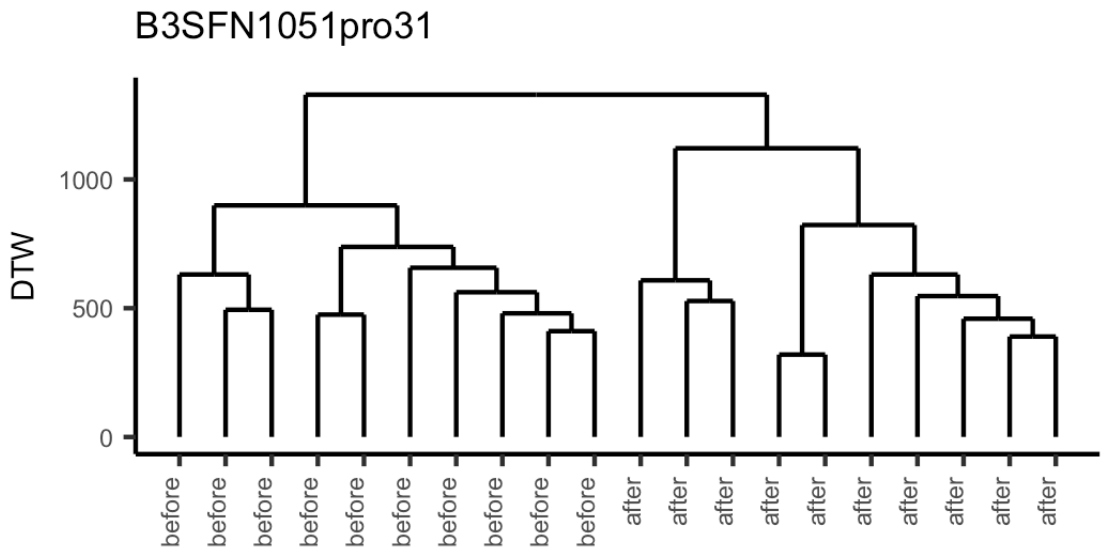

図 5: デンドログラム 2

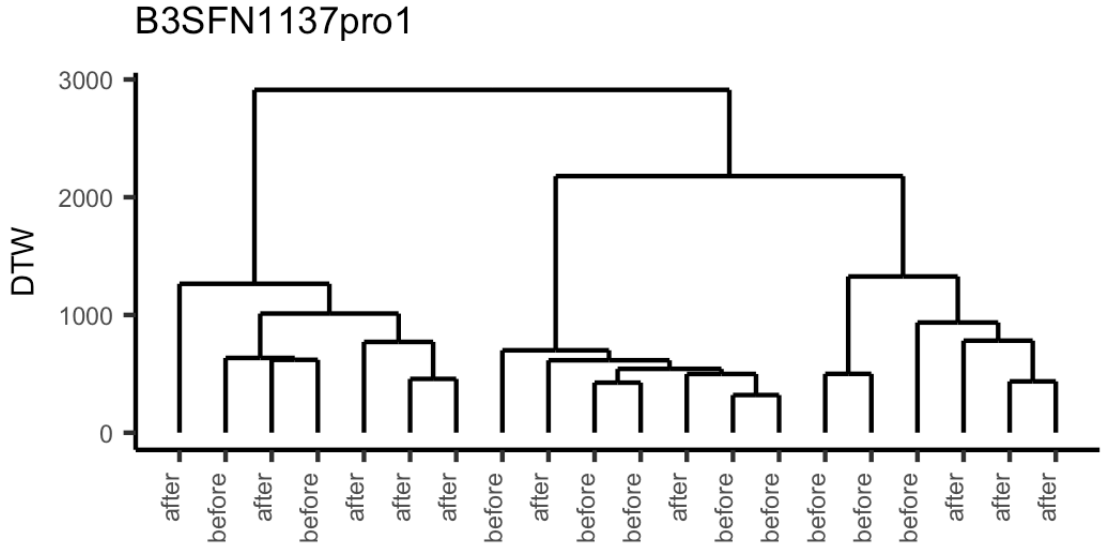

図 6: デンドログラム 3

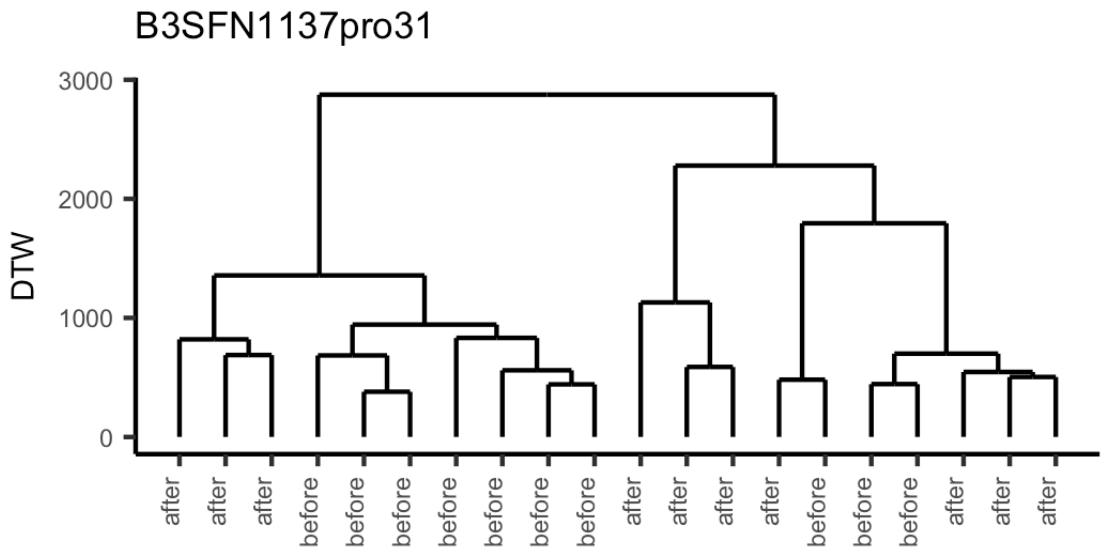

図 7: デンドログラム 4

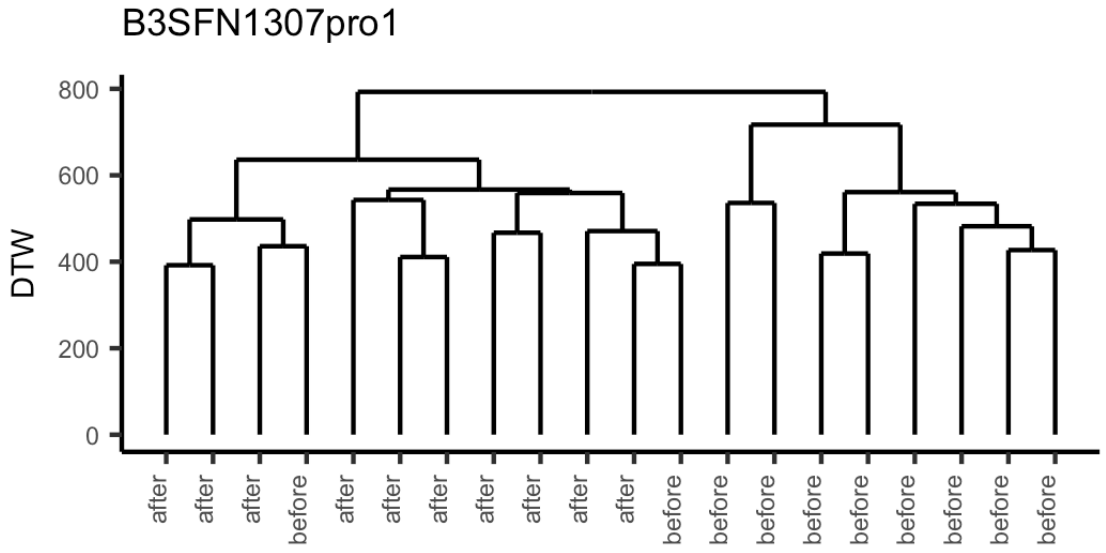

図 8: デンドログラム 5

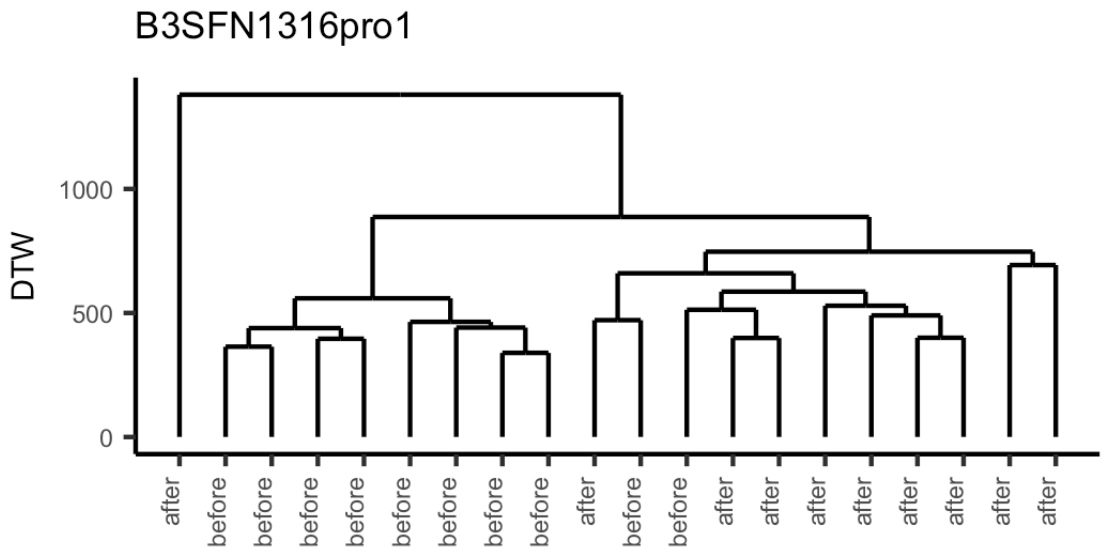

図 9: デンドログラム 6

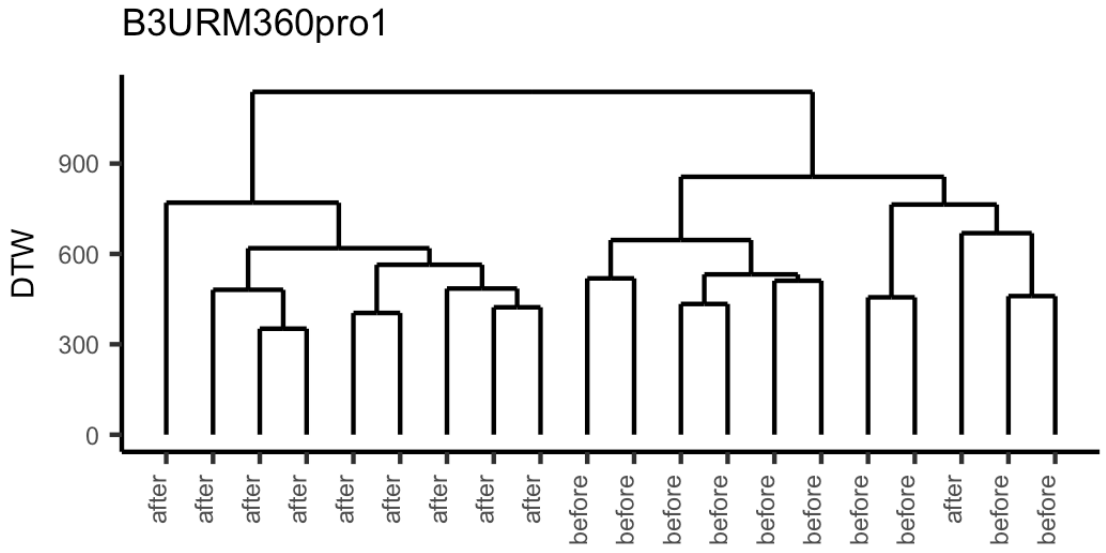

図 10: デンドログラム 7

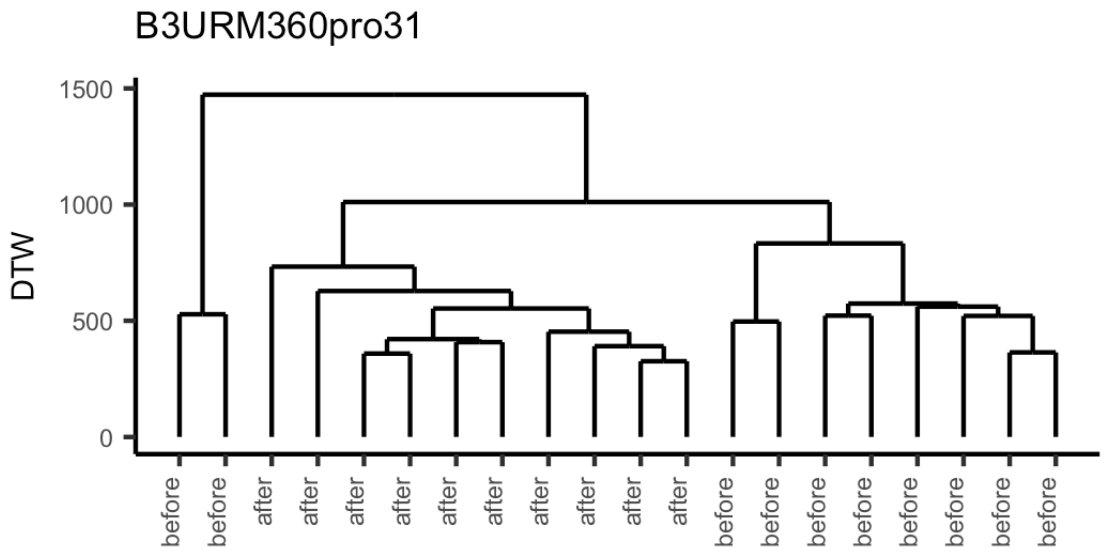

図 11: デンドログラム 8

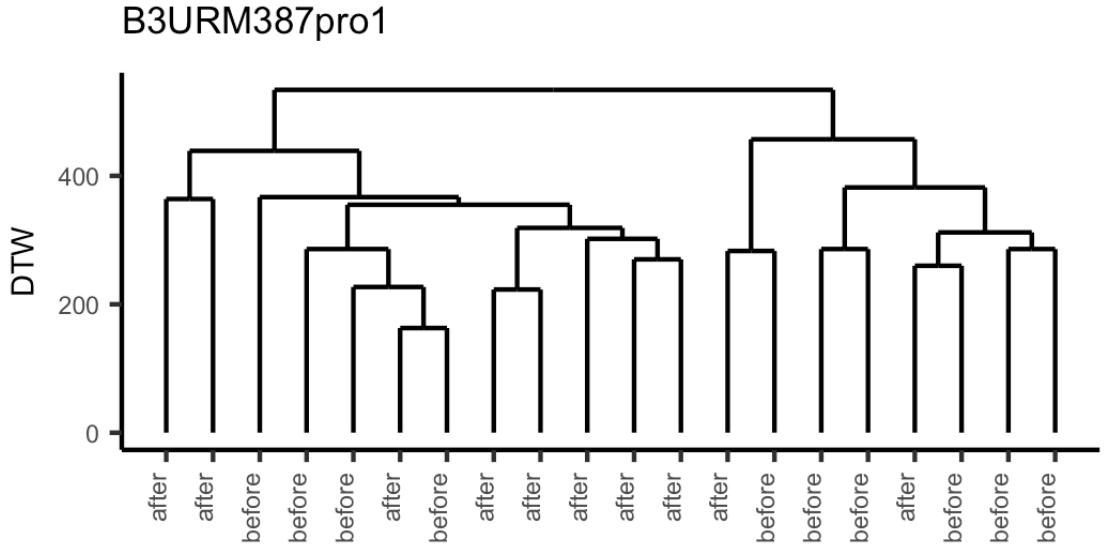

図 12: デンドログラム 9

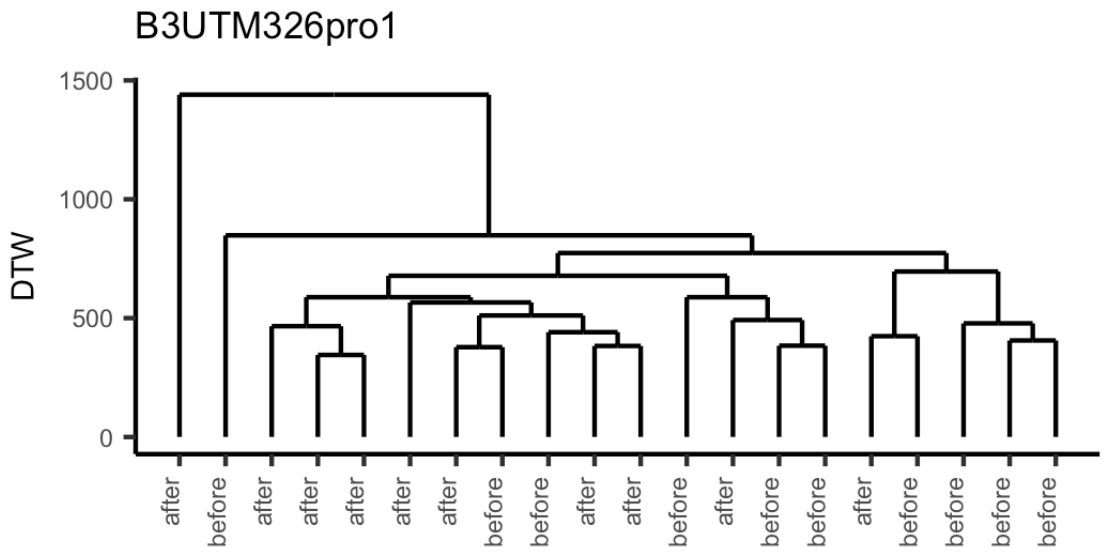

図 13: デンドログラム 10

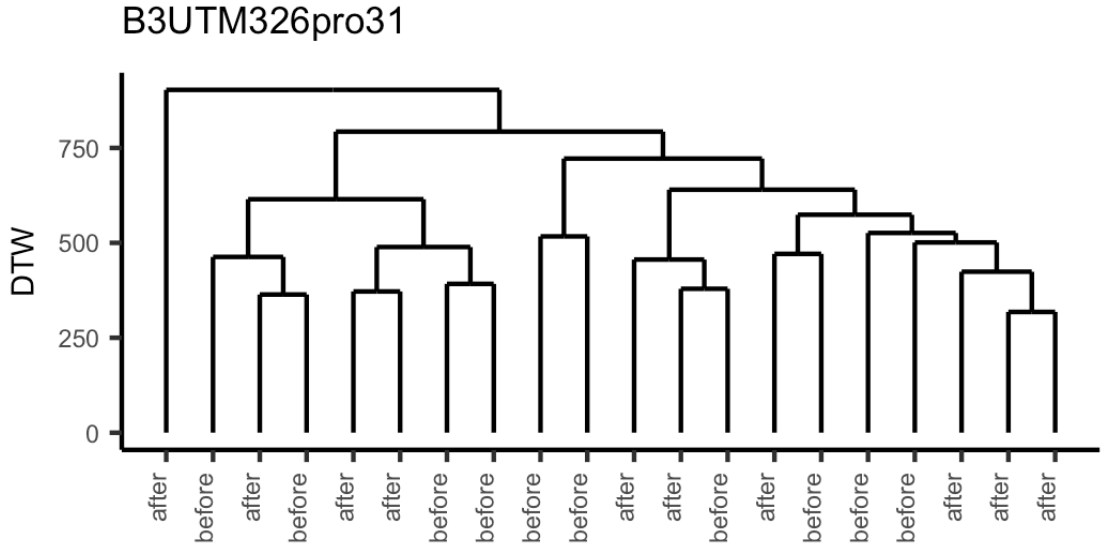

図 14: デンドログラム 11

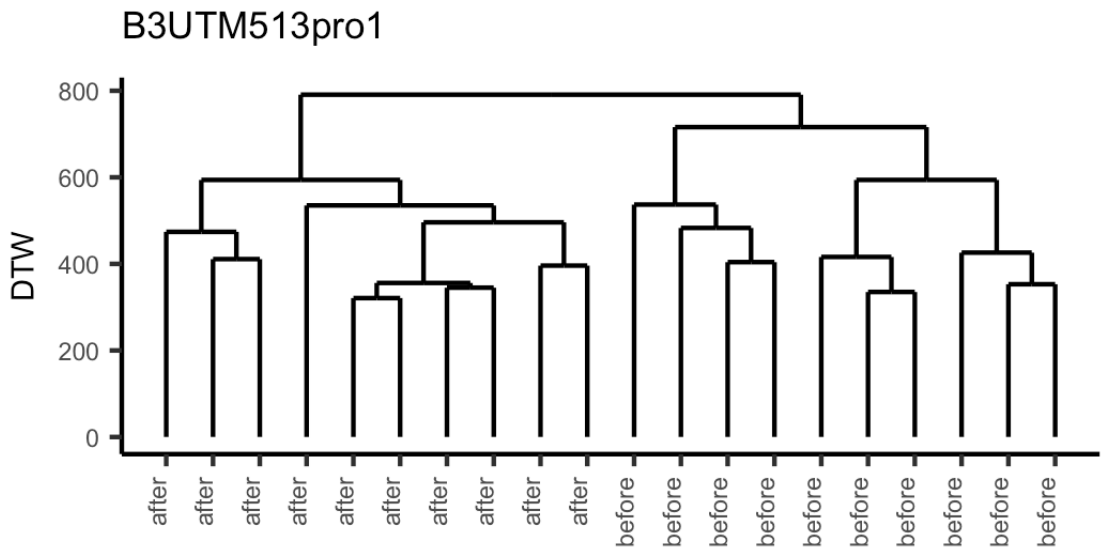

図 15: デンドログラム 12

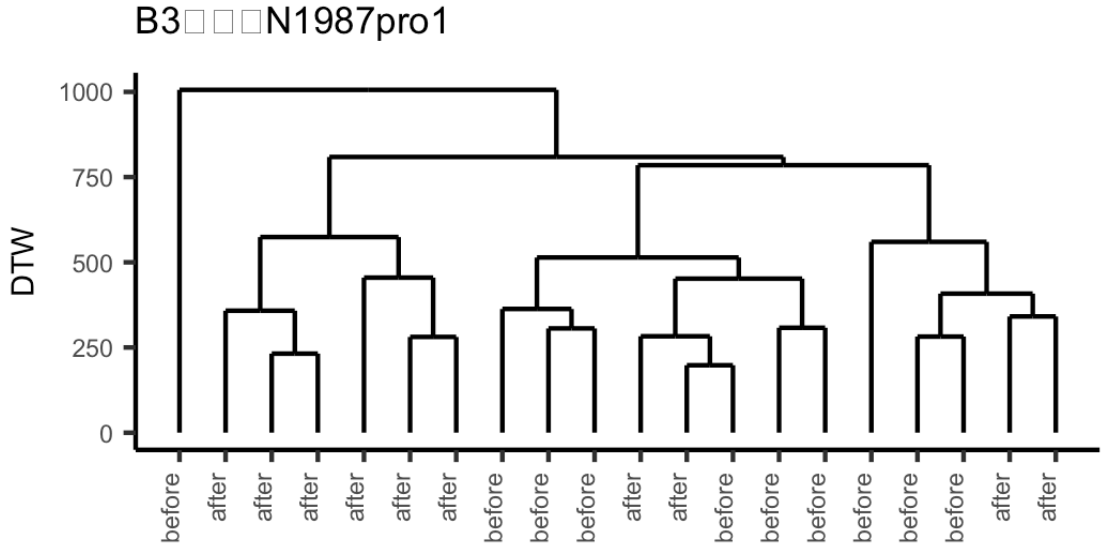

図 16: デンドログラム 13

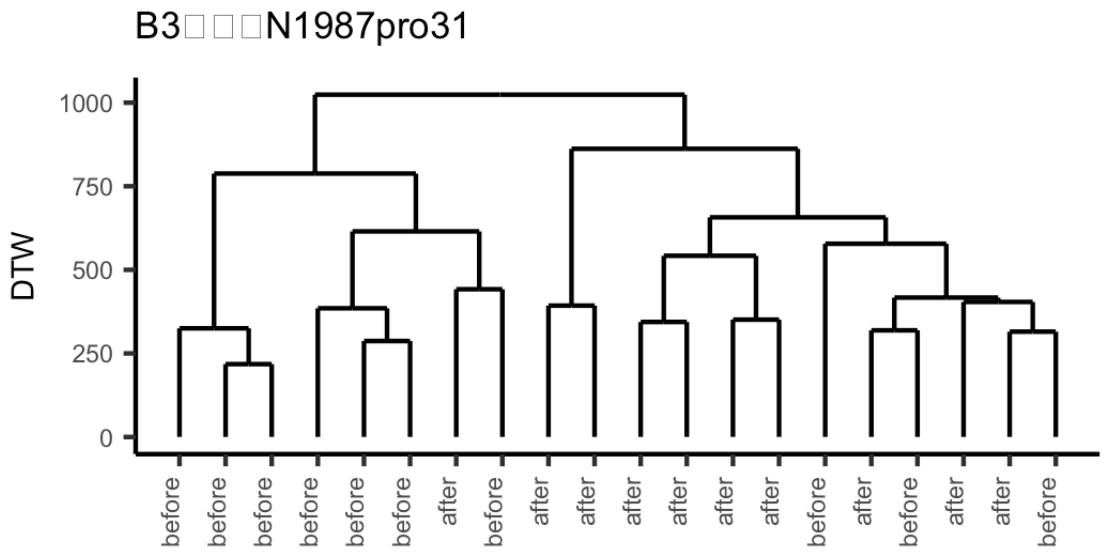

図 17: デンドログラム 14

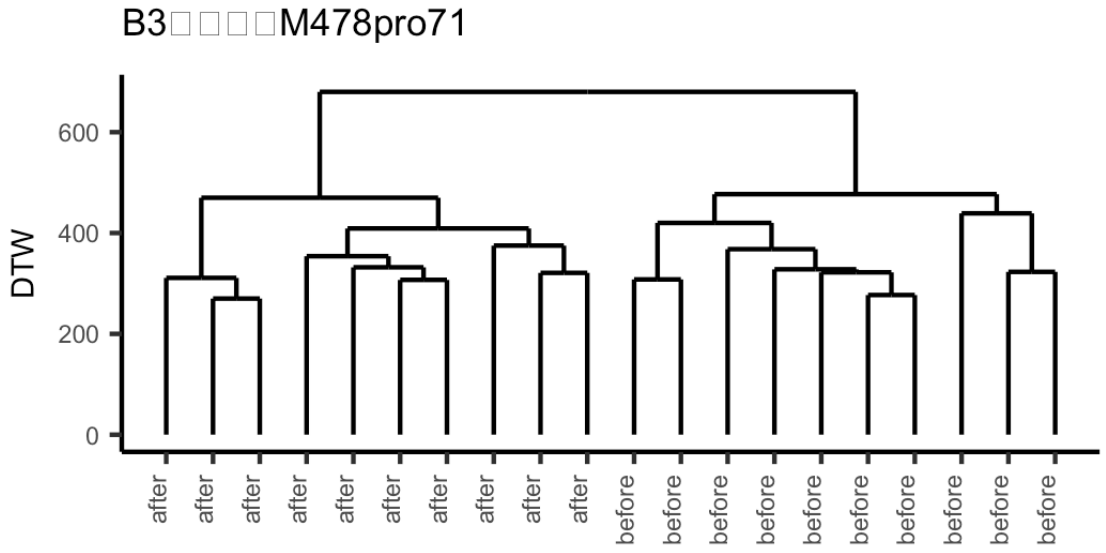

図 18: デンドログラム 15

4　考察・今後の課題

参考文献

[1] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, Vol. 28, No. 4, pp. 511–533, Nov 2002.

[2] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, Vol. 6, No. 1, p. 17, Aug 2017.

[3] Javier Lamar, Edel Garcia, and Rocio Gonzalez-Diaz. Human gait identification using persistent homology. pp. 244–251, 09 2012.

[4] Daniel Platt, Saugata Basu, Pierre Zalloua, and Laxmi Parida. Characterizing redescriptions using persistent homology to isolate genetic pathways contributing to pathogenesis. Vol. 10, , 01 2016.

[5] Marian Gidea. Topological data analysis of critical transitions in financial networks. In Erez Shmueli, Baruch Barzel, and Rami Puzis, editors, *3rd International Winter School and Conference on Network Science*, pp. 47–59, Cham, 2017. Springer International Publishing.

[6] Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, Vol. 491, pp. 820 – 834, 2018.

[7] Donald Berndt and James Clifford. Finding patterns in time series: a dynamic programming approach. pp. 229–248, 02 1996.

[8] Ahlame Douzal Chouakria and Panduranga Naidu Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, Vol. 1, No. 1, pp. 5–21, Mar 2007.

[9] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, Vol. 13, No. 4, pp. 411 – 430, 2000.

[10] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, Vol. 27, No. 12, p. 124007, 2011.

[11] Yuhei Umeda. Time series classification via topological data analysis. 人工知能学会論文誌, Vol. 32, No. 3, pp. 1–12, 2017.

[12] Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, 2007.

[13] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, Vol. 33, No. 2, pp. 249–274, Feb 2005.

[14] Floris Takens. Detecting Strange Attractors in Turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, Vol. 898 of *Lecture Notes in Mathematics*, chapter 21, pp. 366–381. Springer, Berlin, 1981.

[15] Nina Amenta, Dominique Attali, and Olivier Devillers. Complexity of delaunay triangulation for points on lower-dimensional polyhedra. In *SODA*, 2007.

[16] Dominique Attali and Jean-Daniel Boissonnat. Complexity of the delaunay triangulation of points on polyhedral surfaces. *Discrete Comput. Geom.*, Vol. 30, No. 3, pp. 437–452, September 2003.

[17] Patrick Truong. An exploration of topological properties of high-frequency one-dimensional financial time series data using tda. Master's thesis, KTH, Mathematical Statistics, 2017.