



入門講座

インフォマティクス入門ー6

トポロジカルデータ解析による 複雑な顕微画像からの特徴抽出

Feature Extraction from Complex Microscopic Images Using
Topological Data Analysis

赤木和人
Kazuto Akagi

東北大学
材料科学高等研究所
准教授

1 はじめに

実用材料を対象としたマテリアルズ・インフォマティクス (MI) では、目標とする材料特性・性能を実現する組成・プロセス条件を予測することが求められる。そのためには組成・プロセス条件と材料特性・性能とを適切に関連づける必要があるが、プロセス条件を完全に記述するパラメータ空間を設定することは一般的には困難であり、機械学習で提示されたプロセス条件が非現実的なものになることも少なくない。そこで、組成やプロセス条件をかなりの程度反映していると期待されるミクロな構造的な特徴に注目する。構造情報を用いることで組成・プロセス条件と材料特性・性能との関連づけを支援できる可能性があるが、複雑な顕微画像から定量的な特徴抽出を行う手法はまだ確立されていない。そこで、本解説では主にプロセス条件に焦点を当て、パーシステント・ホモロジーと呼ばれる新しい数学的な概念に基づくトポロジカルデータ解析 (TDA) を用いたアプローチを紹介す

る。TDAは複雑な構造データに内在する秩序への気づきを助けるツールであり、幾何学的特徴に基づいた定量的かつ扱いやすい記述子を生成できる。画像データ自体ではなくこの記述子を機械学習に用いることでMIの見通しを良くすることが、本アプローチのねらいである (図1)。

2 白黒二値画像のトポロジカルデータ解析

パーシステントホモロジー (persistent homology) は、原子配置やピクセル画像のような離散点の集合が持つ「データの形」を幾何学的な「穴」の情報で表現する数学的な枠組みである。3次元データには0次、1次、2次の穴があり、それぞれ、つながり、リング、空洞に対応する。「穴」の定義や群の構造、その代数演算についての数学的な詳細は書籍や他記事に譲る¹⁻⁴⁾。ここでは顕微画像が持つ特徴をパーシステントホモロジーを使って調べたいので、白黒二値画像において黒

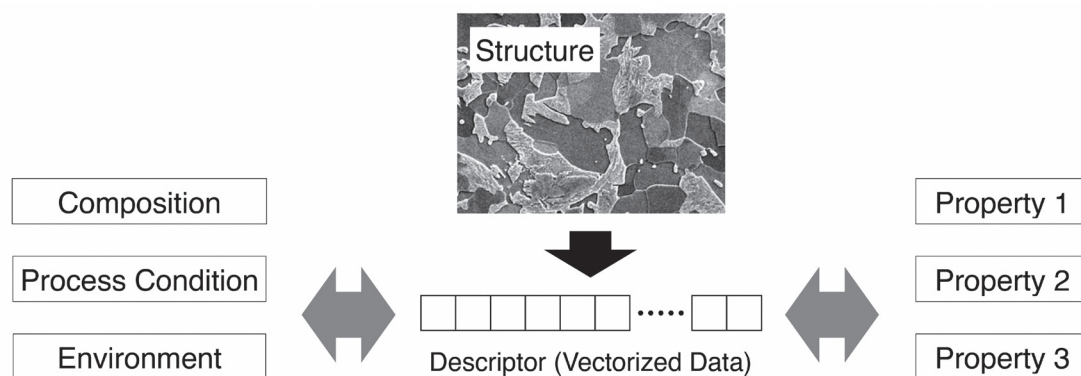


図1 顕微画像が持つ構造の情報を介したマテリアルズ・インフォマティクスの模式図。画像そのものではなく、適切に変換された記述子を介したプロセス条件と材料特性の結びつけを目指す

(もしくは白) の領域を太らせたり細らせたりする操作に伴う穴の生成と消滅を記述することを考える。簡単のため、2次元の場合で説明する。まず、生成と消滅のタイミングを定量的に記すため、例として図2 (a) の黒の領域の表面を基準とした各ピクセルの距離を与える。2つのピクセルの中心間のユークリッド距離を考えても良いが、ここではマンハッタン距離を使うのが便利である。これは、隣り合うピクセルの辺を少なくとも何回横切れば黒のピクセルに到達できるかを数えたもので、図2 (a) にはその数値が記載してある。黒の

領域の内側にも距離は定義でき、それを負の値で表す。黒の領域を1ピクセル太らせる操作は、図2 (b) のようにマンハッタン距離が1の白ピクセルを黒ピクセルにすることに対応する。この手法は整数値のみの扱いで良いため、計算コストや出力データサイズを減らせるメリットがある。

図3 (a) の白黒二値画像 (260px×260px) を用いて、黒の領域を太らせたり細らせたりしながら穴の生成 (birth) と消滅 (death) を調べよう。まず、黒の領域が白の領域を囲むリング構造 (1次のホモロジー) について見ると、図3 (a) です

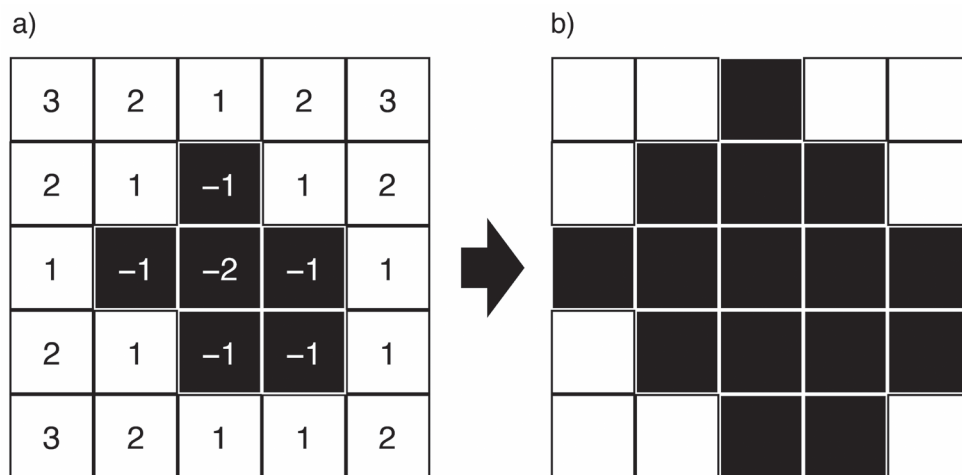


図2 (a) 黒の領域の表面から見た各ピクセルのマンハッタン距離。
(b) マンハッタン距離に基づいて1だけ太らせた黒の領域

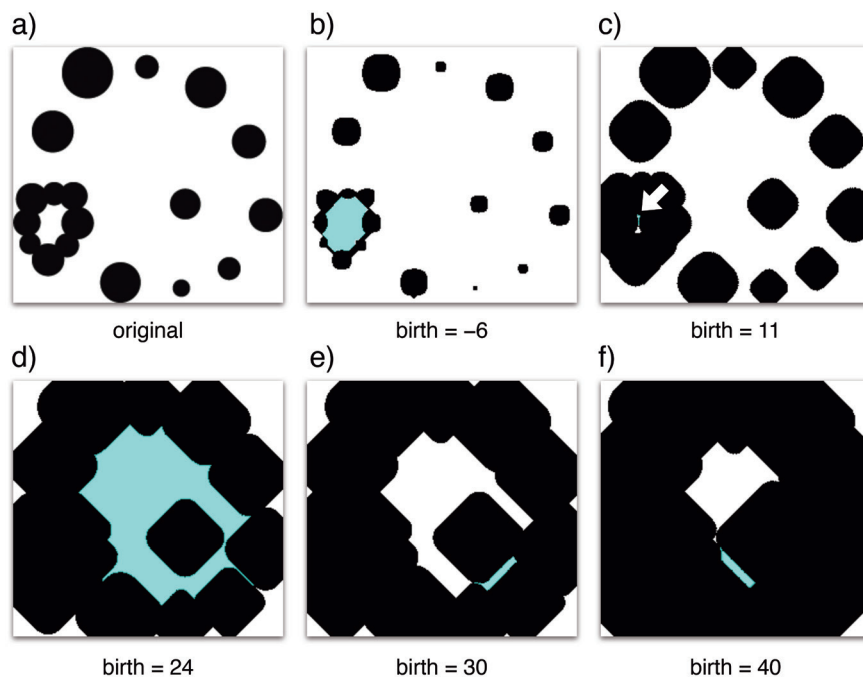


図3 (a) 260px×260pxの元画像。(b) マンハッタン距離に基づいて6だけ細らせた黒の領域。このタイミングで生成したリング構造を水色で示す。(c) -(f) 同様に11, 24, 30, 40だけ太らせた黒の領域と、そのタイミングで生成したリング構造。(c) の穴は小さいため、白抜き矢印で位置を示した

に存在しているリングは黒を7だけ細らせると切れてしまうため、そのbirth値は-6となる。図3 (b) にその時の状態を示す。図3 (a) の状態から黒を11だけ太らせると図3 (c) のように小さなリングが一時的に現れ (birth値は11)、すぐに消滅する (death値は12)。このリングは (birth, death) = (11, 12) として記録される。黒を13だけ太らせると残ったリングも消滅し、図3 (a) に最初からあるリングは (-6, 13) と記録される。同様に、図3 (d) (e) (f) で生成したリングは、それぞれ (24, 75) (30, 35) (40, 45) と記録される。これらを2次元のbirth-deathプロットとしてまとめると、図4 (b) の1次のパーシステント図 (PD, persistence diagram) が得られる。プロット上にある各birth-death点をジェネレータと呼び、(24, 75) のようにbirth値とdeath値の差 (lifetime) が大きなジェネレータは大きく等方性の良いリングに、(11, 12) や (30, 35) のようにlifetimeが小さなジェネレータは小さいまたは扁平なリングに対応する。図3で取り上げていない小さなリングがあと2つ記録されていることも分かるだろう。

一方、図4 (a) に示すPDは0次のホモロジーに対応する。0次のPDに含まれるジェネレータは、birth値が黒い領域の大きさを、death値が黒い領域間の距離を表す。黒の領域は新たに生成することがないため、birth値は必ず負の値になる。正のdeath値を持つグループ①は図3 (a) において黒の領域がつながるイベントを、負のdeath値を持つグループ②は図3 (a) に存在するリングを構成する黒の領域が分裂するイベントを表す。なお、黒の領域について0次と1次のPDが得られたように白の領域についても0次と1次のPDが得られるが、黒の0次と白の1次、黒の1次と白の0次がそれぞれほぼ同じ情報を持つことを指摘しておく。

このように、PDは入力データの構造を一意に定量化した系の「指紋」となる。従来のホモロジーも「穴」に着目して対象を特徴づけることができるが、ドーナツとマグカップを同じものとみなすほど大胆な抽象化を行う。パーシステントホモロジーは大きさや歪みの情報を残した適度な抽象化を行うため、材料構造のコンパクトかつ定量的な記述に使いやすい。画像データからPDへの変換 (順解析) は不可逆的な情報圧縮であるため、PDの情報だけから元の構造を再現することはできない。しかし、PDを計算する際の中間データであるインデックス情報を残しておくことで、各ジェネレータが元の構造のどこに対応しているのかを辿れる (逆解析)。図3 (b) - (f) におけるリング構造の着色はこれを利用している。これらの計算と解析は、無償で公開されているソフトウェア「HomCloud」⁵⁾ を用いて行うことができる。TDAによる画像の取り扱いについては文献6を、顕微画像の具体的な解析例としては文献7も参考にされたい。

蛇足であるが、図3 (c) に示す1次のホモロジーにおけるリング構造 (11, 12) はリング構造 (-6, 13) の生成・消滅の途中に副次的に現れたものであり、(-6, 13) と (11, 12) は親子関係にある。このような情報を用いることでより詳細な解析も可能になる。誌面の都合で詳細は割愛するが、2次元の散布図としてのパーシステント図はパーシステントホモロジー群の情報の一部にすぎないことは覚えておいてほしい。

3 鉄鋼組織画像の解析と定量化

この節では、戦略的イノベーション創造プログラム (SIP) 「統合型材料開発システムによるマテリアル革命」の標準試

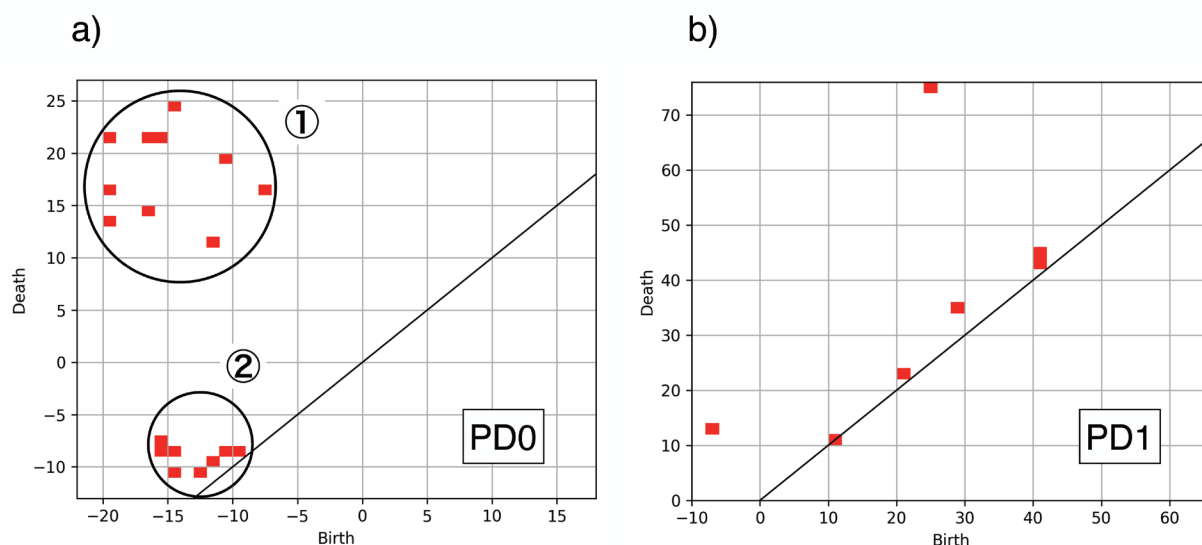


図4 (a) 図3 (a) を特徴づける0次のパーシステント図(PD0)。①は黒の領域を太らせることで繋がるものに、②は細らせることで切れるものに対応する。(b) 1次のパーシステント図(PD1)。birth値が負のものは、元の画像に存在するリング構造に対応する

料として提供されているDP鋼の走査電子顕微鏡 (SEM) 画像を対象とした解析例を紹介する。図5は組成と熱処理条件が異なる9種類の試料の1000倍画像で、圧延面に並行な面を観察したものである。撮影者・撮影機材・撮影条件は同じとなっており、暗い部分がフェライト相、明るい部分がマルテ

ンサイト相に対応している。これらの256階調グレースケール画像が持つ幾何学的特徴を輝度の情報に基づいて抽出することも可能であるが、前処理や解釈における実用上の困難がある。そこで、図6のように白黒二値化した画像を取り扱う。ここでは、提供されたSEM画像 (1280px×960px) を平均輝

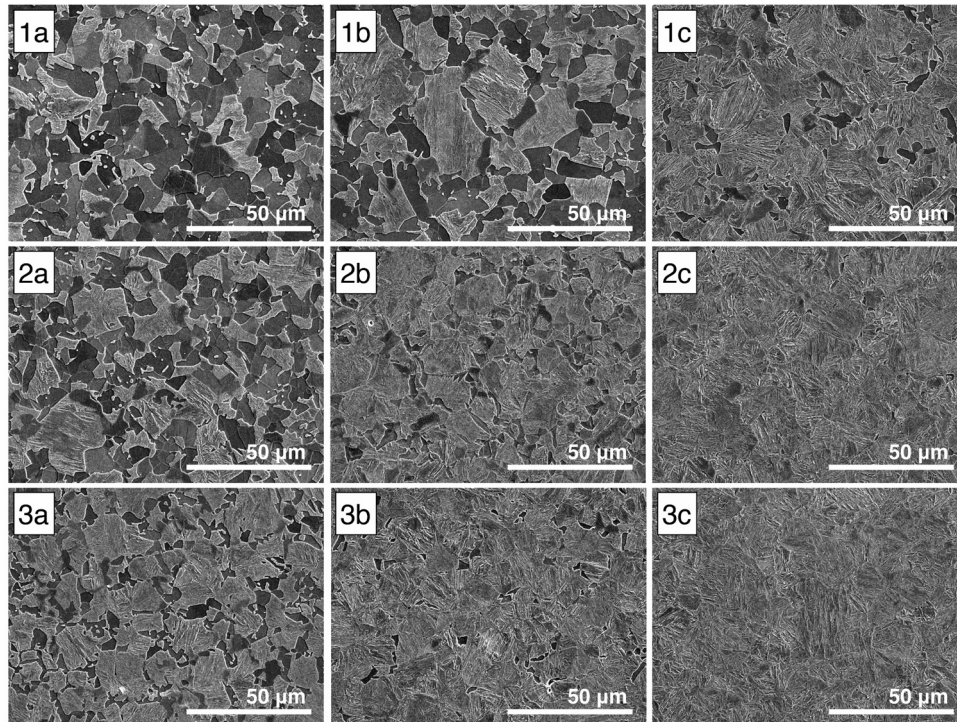


図5 本解説で扱う9種類の試料に対応する走査電子顕微鏡画像(1000倍)の例。1a-3cは試料を識別するためのインデックス

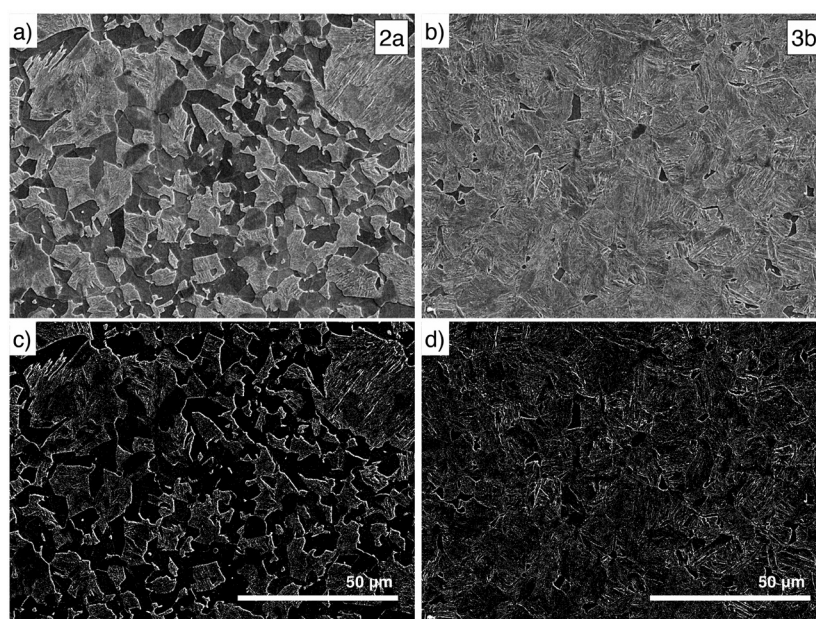


図6 (a) (b) 試料2aおよび3bの顕微画像(1280px×960px)の輝度を規格化したもの。(c) (d) それらを二値化し、3pxのカーネルによる収縮・膨張処理により白の輝点のノイズを低減したもの

度と輝度分布について規格化した後 (図6 (a) (b))、同一の閾値を用いて二値化した (図6 (c) (d))。

図6 (c) (d) の白の領域について得られたPDを図7に示す。色は同じ birth-death 値を持つジェネレータの個数を表している。図7 (a) (b) は0次のホモロジーの情報を表しており、

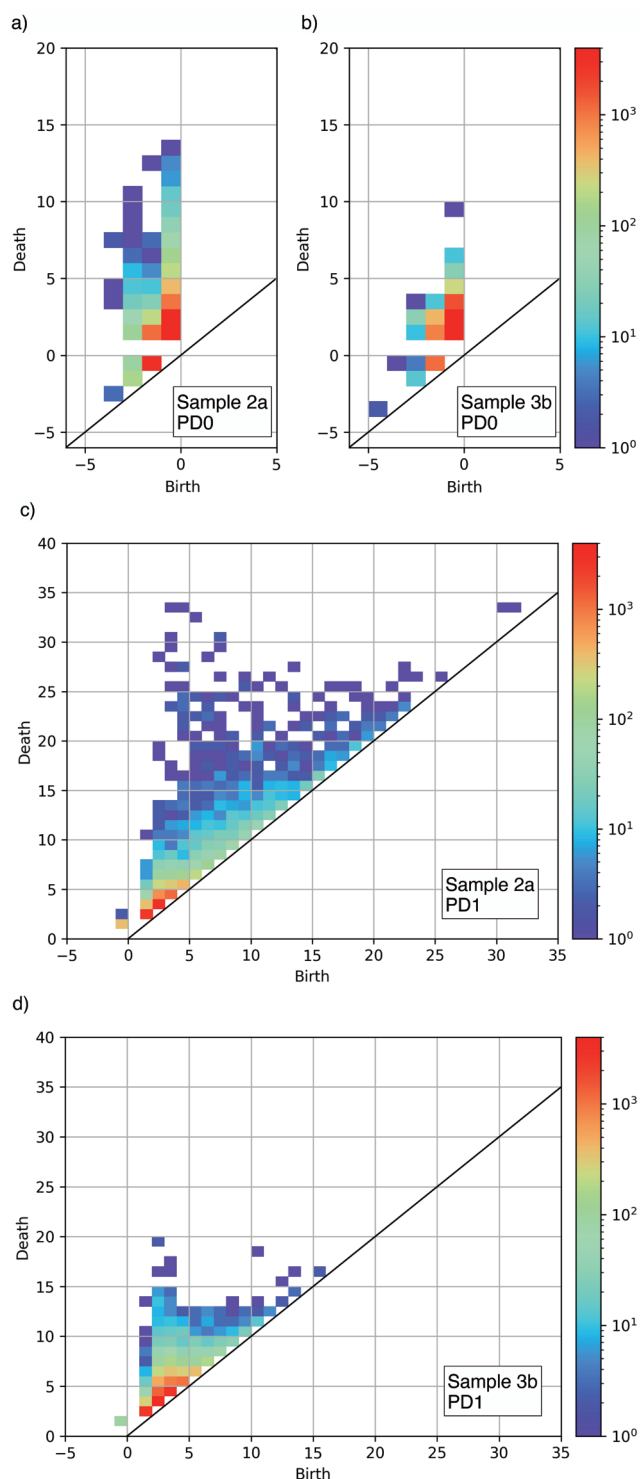


図7 (a) (b) 試料2aおよび3bの白の領域についての0次のパーシステント図(PD0)。色は同じ birth-death 値におけるジェネレータの個数を表す。(c) (d) 試料2aおよび3bの白の領域についての1次のパーシステント図(PD1)

試料2aは試料3bよりも広い範囲にジェネレータが分布している様子が分かる。負の大きな birth 値は白の領域に厚みがあることを示し、正の大きな death 値は白い領域が互いに離れた構造を有していることを示す。これはマルテンサイト相の形や分布についての視覚的な特徴と一致している。同様に、図7 (c) (d) は1次のホモロジーの情報を表しており、やはり試料2aのPDの方が大きく広がっている。lifetime が長いジェネレータの多さは、多数の大きなフェライト相がマルテンサイト相と入り乱れて存在している様子を示している。ジェネレータの数からも分かるように、0次のホモロジーよりも1次のホモロジーの方が情報量が多いため、以降は1次のホモロジーに絞って解析を進める。

PDに捕らえられた鉄鋼組織の形の特徴がどのくらいの定量性を持っているかを調べるために、PDをベクトルに変換し、異なる画像間の類似性をベクトルの内積 (cosine 類似度) として評価することを考える。2次元のヒストグラムであるPDは、birth 値と death 値の範囲を決めて1次元的に並べ直すことで、同じ birth-death 値を持つジェネレータの個数を各要素とするベクトルに変換できる。図7 (c) (d) の1次のPDだけでなく他の試料のPDも踏まえて birth 値と death 値の範囲をいずれも -5 から 65 までと決めることで、対角線の上側のみを対象として2485次元ベクトルが得られる。図5の9種類の試料画像は同じ面に対して100枚ずつ用意されているため、合計900本のベクトルに対して総当たりで内積を評価する。図8 (a) は試料1aを基準として内積値の分布を見たものである。同じ試料1aの異なる場所を撮影した100枚の画像から生成したベクトルは、互いの内積値が0.95前後の狭い範囲に収まっている。一方、試料1aから生成したベクトル群と試料1b, 1cから生成した各ベクトル群との内積値は、それぞれ0.88, 0.80にピークを持つ別のグループを形成している。これは、試料1a, 1b, 1cがそれぞれ異なるグループとして識別されたこと、試料1aから見ると試料1cよりも試料1bの方が類似度が大きいと評価されたことを示す。比較のために、各画像の4分の1の領域 (640px×480px) だけを用いてPDを計算して900本のベクトルを生成すると、図8 (b) のような結果が得られた。ベクトル長は同じく2485である。試料1aから生成したベクトル間の内積値の分布が広がるだけでなく、試料1aから見て試料1bと試料1cの区別がつかなくなっている様子が分かる。このように、生成した記述子ベクトルの品質は画像データが各試料の構造的特徴を十分に含むかどうかによって依存する。

画像に含まれる幾何学的特徴をひとたびベクトル化してしまえば、種々の機械学習の手法を容易に適用できる。ここでは、定量性があり解釈が容易な主成分分析を用いることにする。主成分分析を行えば、与えられたベクトルのセット $\{v_1, v_2, \dots, v_n\}$

に対して、それらの違いが際立つ軸（主成分ベクトル v_{PC1} , $v_{PC2}, \dots$ ）を見つけることができる。 $\{v_1, v_2, \dots, v_n\}$ の平均ベクトルを v_{avg} とすれば、ベクトル v_i の第1主成分値は $(v_i - v_{avg}) \cdot v_{PC1}$ で与えられる。「 $\cdot$ 」は内積を表す。 v_{PC1} , v_{PC2} , $\dots$ と v_{avg} を学習に用いたベクトルセットから生成した「ものさし」だと思えば、別の新たな試料 v_{new} を持ってきて $(v_{new} - v_{avg}) \cdot v_{PC1}$ のように測ることもできる。

用意された画像データのセットに共通の前処理（輝度の規

格化と白黒二値化）を施し、白い領域についての1次のPDから生成したベクトルに対する主成分分析を行なった。試料1aから3cの900本のベクトル（撮影方向は同じ）の第1主成分と第2主成分について2次元プロットしたものが図9 (a) である。第1主成分と第2主成分の寄与率は、0.90と0.08であった。試料ごとの100個の点はそれぞれ島状のグループを形成しながら分類されている様子が分かる。試料3bは島が他と比べて広がっていることから、構造の不均一性が大きい

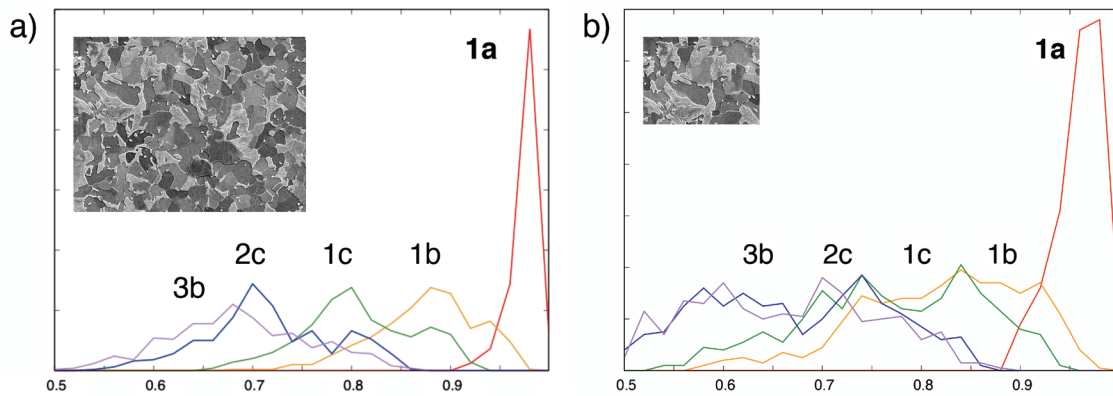


図8 (a) 1280px×960pxのサイズの画像の白の領域についての1次のパーシステント図をベクトル化し、試料1aに属する100個のベクトル同士の内積値、および試料1aから見た他の試料のベクトルとの内積値の分布を記したものの。(b) 640px×480pxのサイズの画像に対して同様の分布を記したものの

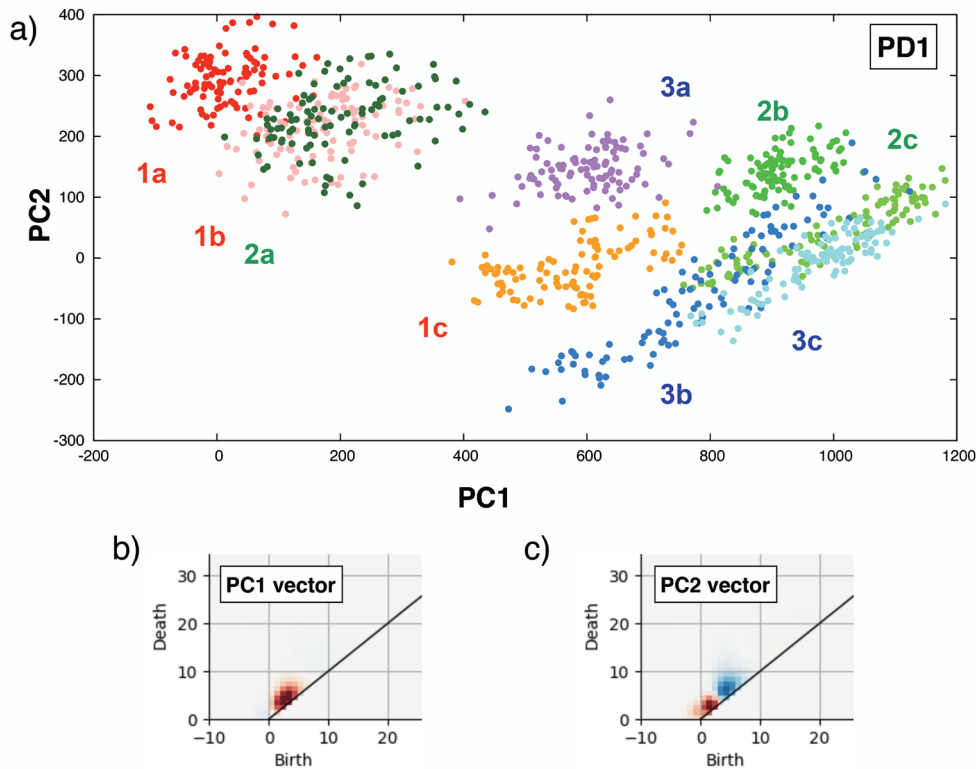


図9 (a) 試料1a-3cの白の領域についての1次のパーシステント図に対応する900個のベクトルを、第1主成分(PC1)と第2主成分(PC2)を用いて2次元プロットしたもの。(b) パーシステント図と同じインデックスを用いて第1主成分ベクトルを可視化したもの。(c) 同じく、第2主成分ベクトルを可視化したもの。赤は正の要素を、青は負の要素を表す

ことが示唆される。この分類はPDの差異を定量化して得られたものであるが、PDのどこに違いを見出したのかを知るためには主成分ベクトルを見ればよい。図9 (b) (c) は第1主成分ベクトル v_{PC1} と第2主成分ベクトル v_{PC2} の各要素をPDと同じインデックスで2次元に並べたもので、正の値を持つ要素を赤く、負の値を持つ要素を青く表示している。図7 (c) (d) などと照らし合わせてみると、かなり小さなリング構造の情報を手掛かりに試料が分類されていることが示唆される。まだ解析途上であるため本稿では詳細を割愛するが、得られた主成分と引張強度の測定値が良い相関を示しているのは興味深いところである。なお、白の領域の0次のホモロジーを用いた主成分分析では第1主成分が0.96の割合を占め、第2主成分は試料の分類に寄与しなかった。この0次のホモロジーの第1主成分による分類は1次のホモロジーの第1主成分による分類と非常に似たものであった。

4 考察とまとめ

PDを構成するジェネレータは、そのlifetimeに応じて階層的に異なる大きさの構造を捉えている。例としていくつかのlifetimeに対応する構造的特徴を逆解析によって取り出し、試料1bの白黒二値化画像に重ねて着色した (図10)。前述のように、図9 (b) (c) の主成分ベクトルはlifetimeが5程度まで、かつbirth値が小さいリング構造に基づいて試料を分類している。詳細に逆解析をしてみると第1主成分はマルテン

サイト相の内部を見ているのに対し、第2主成分はマルテンサイト相の表面近くを見ていることがわかった。我々が主観的に形と感ずる海島構造とは異なる情報を用いて試料の分類や引張強度との相関が取れていることは興味深い。一方、我々の目で形として感ずるパターンはlifetimeが7あたりから出現し始めるが、試しにlifetimeが5以下のジェネレータを取り除いて主成分分析を行ったところ、試料の分類性能は著しく低下して主成分と引張強度との相関も失われた。大きなスケールの構造情報が試料の分類や材料特性との相関に寄与するかどうかの結論を得るには低倍率で広い視野を撮影した画像データ (これらの構造を十分に含むもの) を用いた解析が必要である。

lifetimeが小さな構造は、白黒二値化やノイズ除去に左右される脆弱な情報でもある。そこで、今回使用した画像が同一条件で撮影されたものである点を活かし、グレースケール画像についての解析も行った。元のグレースケール画像の輝度情報を規格化後、レベルセット法を適用して計算した1次のホモロジー (暗い領域が明るい領域を囲むリング構造を実空間ではなく輝度空間で調べたもの) をベクトル化した場合でも、図11のように第1主成分について類似の分類を得ることができる。もちろん、撮影者や撮影機材が異なれば画像の陰影は大きく変わるし、どの方向の面を見るかで分類の良否も変わる。実際のMIに用いるにあたっては、データ格付けとその利活用のため、顕微画像を撮影する際の要件と二値化画像を生成する際の手続きをある程度決めておく必要がある。

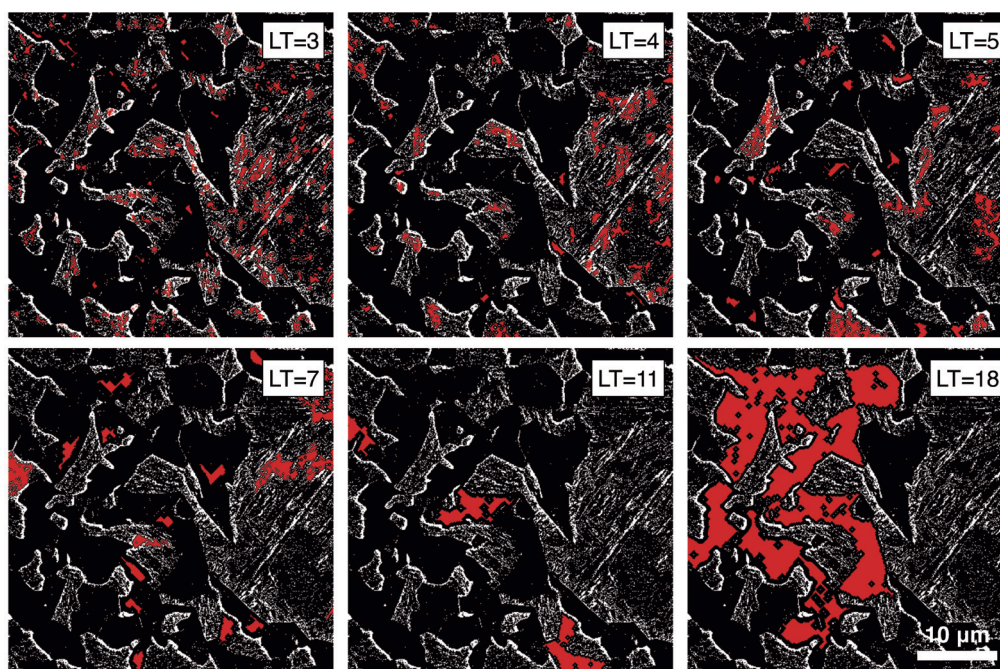


図10 試料1bの白の領域についての1次のパーシステント図を構成するジェネレータに対応する領域を、6つの異なるlifetime (LT) について元画像上に表示したもの

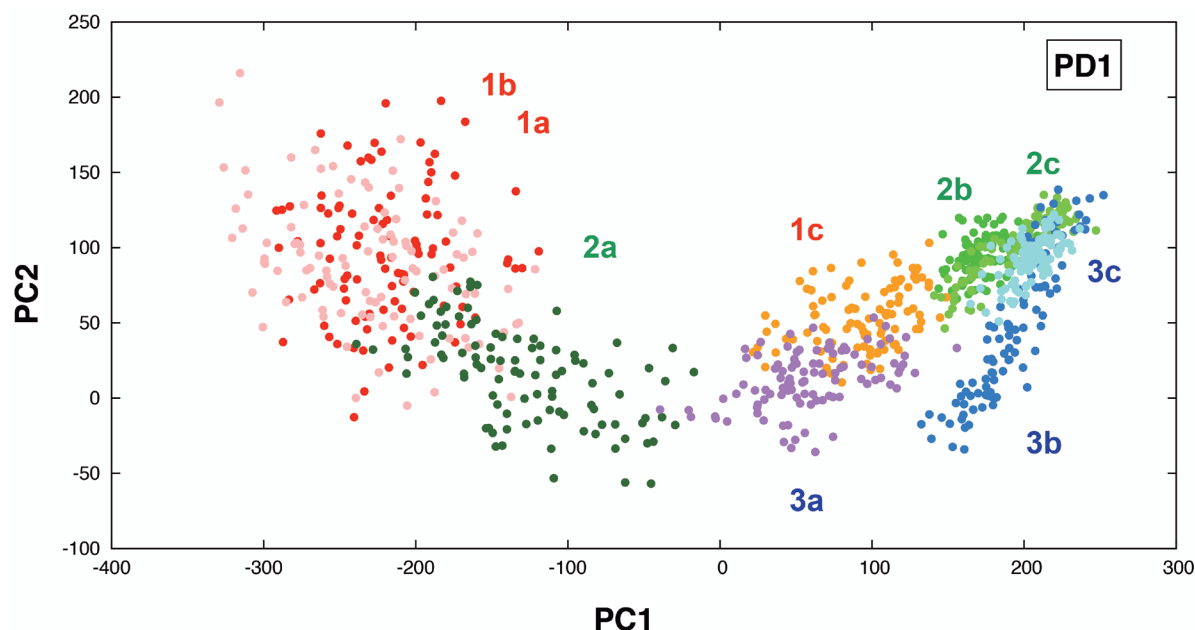


図11 試料1a-3cのグレースケール画像のsublevelsetについての1次のパーシステント図に対応する900個のベクトルを、第1主成分(PC1)と第2主成分(PC2)を用いて2次元プロットしたもの

る。その設計はSIP第2期の後半の課題であり、取り組みの成果については別途報告の機会を持ちたい。

本解説では、パーシステントホモロジーという数学の手法を用いて顕微画像データがどのように定量化できるかを見てきた。さらに定量化した情報(ベクトル化したPD)を記述子として、同一試料内の構造の不均一性を調べたり、試料画像の分類を行ったりできることを紹介した。グレースケール画像から白黒二値化画像への変換としては他のアプローチも考えられ、例えば輝度の微分をエッジとする白黒二値化をしても良い。パーシステント図は画像の向きに依らない情報であるが、結晶の面方位の情報を付加した多次元データを解析することもできる。重要なのは、物理的な妥当性が期待される構造-特性相関や構造-プロセス相関を示す構造的特徴を捉えることである。まだ調べるべき課題は多く残っているが、パーシステントホモロジーに基づくトポロジカルデータ解析を、金属の組織構造に関する現場の経験知や暗黙知をMIに取り入れる手法のひとつとすべく発展させたい。

謝辞

本解説記事は、JST/戦略的イノベーション創造プログラム(SIP)の第2期課題「統合型材料開発システムによるマテ

リアル革命」における研究開発成果に基づくものである。また、パーシステントホモロジーに基づくトポロジカルデータ解析を行うためのソフトウェア基盤「HomCloud」の開発者である大林一平・平岡裕章の両氏への特段の謝意を表する。

参考文献

- 1) 平岡裕章：タンパク質構造とトポロジー：パーシステントホモロジー群入門，共立出版，(2013)
- 2) 平岡裕章，大林一平，赤木和人：人工知能，34 (2019)，3，330.
- 3) 平岡裕章：数理科学 53，サイエンス社，(2015)，48.
- 4) 平岡裕章，西浦廉政：日本物理学会誌，72 (2017)，632.
- 5) I.Obayashi, Y.Hiraoka and M.Kimura：J. Appl. Comput. Topology, 1 (2018)，421.
- 6) HomCloud：https://homcloud.dev/, (2020)
- 7) M. Kimura, I. Obayashi, Y. Takeichi, R. Murao and Y. Hiraoka：Sci. Rep., 8 (2018)，9.

(2020年12月2日受付)