卒業論文
機械学習による質量分析イメージングデータの特徴量抽出

松本　涼太
大阪大学理学部生物科学科生命理学コース
豊田研究室 B4

2024 年 2 月 19 日

目次

概要

質量分析イメージング (mass spectrometry imaging; MSI) は, 生体組織切片の数千点におよぶ計測領域の質量分析を行い得られる, 膨大なマススペクトル情報から, 任意のイオン情報を選択することで, 試料成分の二次元的分布強度を画像化することが可能である. 一方, 分子情報は数百から数千に及ぶため, 注目する組織に多く分布する分子情報を, 効率的かつ客観的に選択することが難しい.

本研究では, 近年急速に発展する機械学習の手法を用いて, マウス精巣組織の多次元大容量データを解析する方法を検討した. その結果, 教師あり学習である線形判別分析および線形 SVM の手法を利用することにより, 野生型マウス精巣組織と, 脂質の生合成に関与する特定の酵素をノックアウトしたマウスの精巣組織を高精度で判別できた. さらに, 学習によって得られた判別関数を用いることで, 脂質を選別することができた. 本研究の結果は, 機械学習を用いることで, 多次元 MSI データの解析は疾患に伴う生体組織の脂質変化を捉えるために有効となりうることを示している。

1　はじめに

WHO の統計では, 不妊症の 48 %は男性側が関与するものであり, 男性不妊の原因として, 精子の形態異常が考えられる. 特に, 精子頭部の異常は精巣内での精子の成熟に関係があると言われている. 哺乳類の精子は, 精巣内にある迂曲した管, 曲精細管 (convoluted seminiferous tubule ; CST) で産生される. 精粗細胞から一連の過程を経て精子になるまでを精子発生 (spermatogenesis) と呼び, その中でも最後の段階である精子細胞が精子になる過程を精子形成 (spermiogenesis) と呼ぶ. 曲精細管の断面図では, 外側から内部へと従って, 精子発生の段階が進み, 成熟した精子は精細管を通って輸送される. 精子形成に異常が生じる場合, この曲精細管内のいずれかの部分で問題が発生していると考えられる.

質量分析イメージング (MSI) は, 生体試料中の化学成分の分布を可視化する方法である. 精子は精巣内の CST で段階的に成熟していくため, MSI によって精巣内の脂質分布を得ることで, 正常な精巣と異常な精巣を比較し, 精子形成のどの段階で違いが生じるかを知ることができると期待される. しかし, MSI によって得られる分子情報は数百から数千に及ぶため, 全データについて, 正常組織と異常組織とを目視で比較することは難しい. また, 正常組織と異常組織とを比較する際に, 明確な違いがあるか否かを判断するには人間の主観が入るため, 客観性に欠けるという問題がある. 多次元のデータに対し, 近年急速に発展する機械学習による解析を適用することで, 多次元のデータから, 正常組織と異常組織でスペクトル強度に有意な差があるものを判別し, 客観性のあるデータを比較できるようになると期待される.

本研究では, 線形判別分析 (LDA) および線形 SVM の手法を利用した. これは, 正常組織と疾患組織の分類にあたってどの変数 (本研究においては質量数／電荷数の比, m/z と表す) が影響を与えているかを示す, 判別関数の重みを得ることができるからである. これらの分析法を用いた結果, 特に線形判別分析で導かれた, 重みの大きい m/z について, イオン像の比較したところ, 野生型とノックアウト型で, CST 内の脂質の信号強度について有意な差があることが確認できた.

2　質量分析

質量分析 (mass spectrometry) [1] とは質量分析装置を用いてイオンの m/z(イオンの質量 m, 電荷数 z, m/z は無次元量) とその強度を測定することである. 原子や分子は質量が非常に小さいため, 重力を用いて計測することが事実上困難なため, 電磁気力を利用して測定する. そのためにはまずイオンを電磁場中で運動させ, 分離することが必要である. 図 1 に, 試料を質量分析する際の流れを示す. 何らかの方法で試料成分をイオン化し, それぞれのイオンを分離後, 質量分析計へと導入され, 各イオンの m/z と信号強度の関係を示す, マススペクトルが得られる. 質量分析を行うことで, 原子の組成比, 化合物の構造解析といった化合物の同定, タンパク質・ペプチドのアミノ酸配列やタンパク質の立体構造といった分子の構造を解析することができる. また, 自然界や生体内に存在する化合物の定量分析, 医薬品や代謝物の体内での分布状況の解析といった, 化学や医学などにも有用な分析法である.

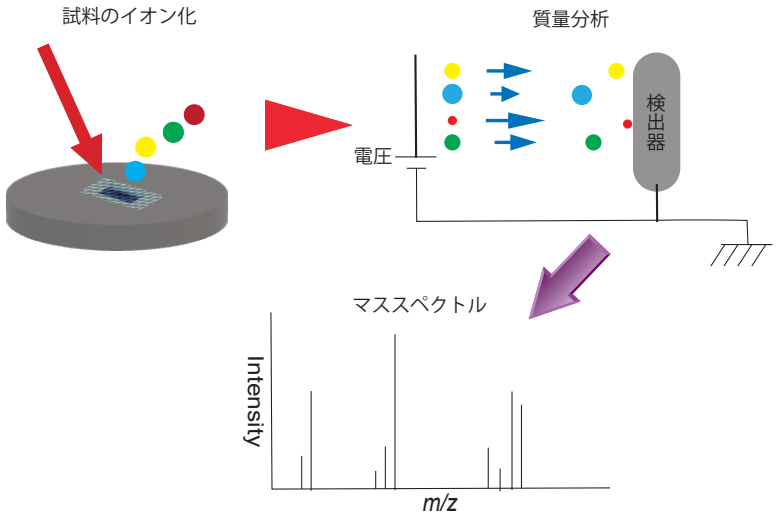


図 1: 質量分析までの流れ

2.1　エレクトロスプレーイオン化法

エレクトロスプレーイオン化法 (electrospray ionization; ESI) は, 溶液試料に, 大気圧下で静電場を印加することで帯電液滴を形成し, それを噴霧しイオン化する方法である. 図 2 に, エレクトロスプレーイオン化の模式図を示す. 高い電圧を印加した emitter から試料溶液を微小な

液滴として噴霧し, その溶媒を蒸発させることでイオンを生成する. ESI ではしばしば多価イオンが生成し, 低い m/z 領域で高分子量測定が可能である. ESI では溶液中に存在するイオンが, 気相に移る機構により, 帯電液滴からイオンが生成する. 帯電液滴からイオンが生成する過程の仕組みについては, 電荷残留モデルとイオン蒸発モデルという 2 つのメカニズムが同時に働くことが考えられている. 両者とも, 帯電液滴が溶媒の蒸発と表面電化の反発により分裂を繰り返す過程は同一であるが, 前者（電荷残留モデル）では, 分裂の結果, 試料のイオン 1 個のみが残る, というモデルであるのに対し, 後者（イオン蒸発モデル）では, 10nm 程度の大きさまで分裂後, 液滴中のイオンが蒸発する, という点が異なる.[1] ESI では分子量範囲が約 1,000,000 以下程度と広範囲で, 液相から気相に直接イオンを取り出すため, 試料分子を分解することなく検出できる, 非常にソフトなイオン化となっている.

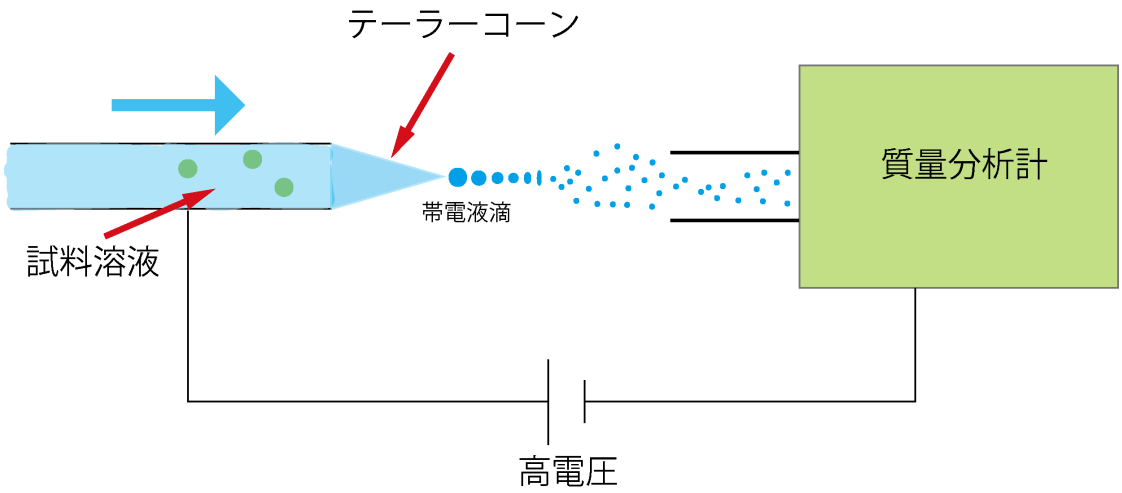


図 2: ESI の方法

2.2 タッピングモード走査型プローブエレクトロスプレーイオン化法

タッピングモード走査型プローブエレクトロスプレーイオン化法 (tapping-mode scanning probe electrospray ionization; t-SPESI) [2] は, 原子間力顕微鏡と, 先ほど述べた ESI を融合したもので, キャピラリープローブを流れる溶媒に電圧をかけ, プローブを振動させながら試料表面とプローブ先端を接触させることで, プローブ先端から流れる溶媒が液架橋 (liquid bridge) を形成し, 試料成分を抽出, イオン化する方法である (図 3). t-SPESI の特徴として, 抽出する溶媒の体積をピコリットル以下にできる点, 短時間で抽出と ESI を実施できる点が挙げられる. 帯電液滴の体積を減少させることで, イオンの検出感度を増加させることができるため, 高空間分解能と検出感度を両立することができる. マトリックスと試料を混合したものにレーザー光を照射させるイオン化法である, マトリックス支援レーザー脱離イオン化 (matrix-assisted laser desorption ionication; MALDI) など, 従来から生体分子の質量分析に使用されてきたイオ

ン化技術と比較し, t-SPESI は, サンプルの前処理を行う必要がなく, また, 試料に大きなダメージを与えることなくイオン化することが可能である点で優れている. 加えて, 計測中に試料ステージを移動させていくことで, 試料中の計測位置を変化させることができるため, マススペクトルとサンプル上での位置情報を同時に得ることができる. これは, 質量分析イメージングを行う上で重要である.

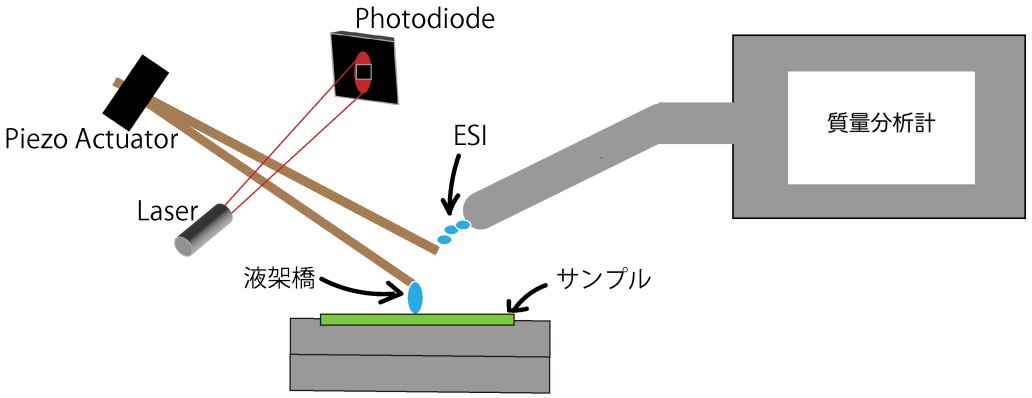


図 3: t-SPESI の模式図

2.3 質量分析イメージング

質量分析イメージング (mass spectrometry imaging; MSI) は, 質量分析により得られる, マススペクトルと, 試料組織の位置情報とを組み合わせ, 試料中の特定の m/z の強度分布を二次元イメージで示す技術である [3]. 図 4 に, MSI の概要を示す. まず試料の微小領域をイオン化し, 質量分析を行った後, 各領域のマススペクトルと位置情報から, 試料中の任意イオンの二次元的強度分布を得ることができる.

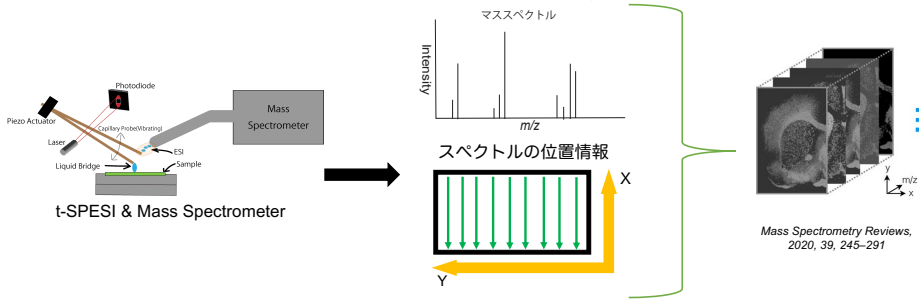


図 4: MSI の流れ

3　脂質とマウス精巣組織

細胞内のリン脂質分布については, ラットの肝臓を用いて詳しく調べられている [4]. その結果から, 以下の複数の事実がわかっている.

- ホスファチジルコリン (phosphatidylcholine; PC) はすべてのオルガネラで最も多いリン脂質である.
- PC は核膜・小胞体＞ゴルジ体＞形質膜と小胞輸送経路にしたがって含有量が少ない.
- ホスファチジルエタノールアミン (phosphatidylethanolamine; PE) とホスファチジルイノシトール (phosphatidylinositol: PI) はすべてのオルガネラにほぼ均等に分布している.
- スフィンゴミエリン (sphingomyelin; SM) はゴルジ体, 形質膜, リソソームに多い.
- ミトコンドリア内膜はホスファチジルセリン (phosphatidylserine; PS) 含有量が低い.
- ガルジオリピン (cardiolipin; CL) はミトコンドリアに局在している.
- リン脂質の合成酵素であるホスファチジルグリセロールホスフェートシンターゼ, ホスファチジルグリセロールホスファターゼ, カルジオリピンシンターゼはミトコンドリア内膜に局在しており, モノアシルグリセロールリン酸がリソソームに局在している.

哺乳類の精母細胞や円形精細胞は, おもに乳酸をエネルギー源として蛋白合成や RNA 合成, 脂質の合成を行っている.

3.1　生体内のリン脂質

脂質とは, 一部例外を除き, 一般に, 水に溶けにくく, 有機溶媒に溶けるものと定義される. 中でもリン脂質は生体膜を構成する主要な成分であり, 膜リン脂質第二位には多価不飽和脂肪酸がエステル結合している. ホスホリパーゼ A2 はこのエステル結合を切断し, リン脂質は, 脂肪酸とリゾリン脂質へ分解される. リン脂質は, 図 5 に示すグリセロリン脂質 (Glycerophospholipids), スフィンゴリン脂質 (Sphingolipids) の 2 つに大別される.

ホスファチジルコリン (phosphatidylcholine)

ホスファチジルコリン (PC) は, 1846 年にはじめて卵黄と脳より単離され, 動物や植物に多く存在するリン脂質である. 特に, 動植物の膜リン脂質においては, その 50 %近くを PC が占めている. その化学構造を図 6 に示す.

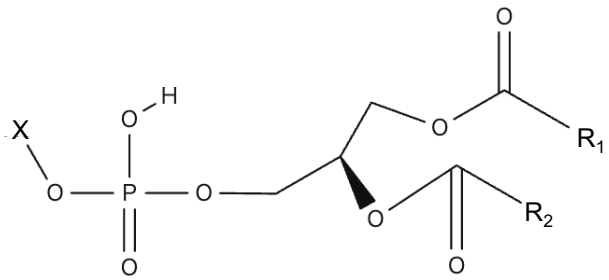


図 5: グリセロリン脂質の化学構造. ホスホリパーゼ A2 によって R_2 の脂肪酸が解離し, リゾリン脂質となる.

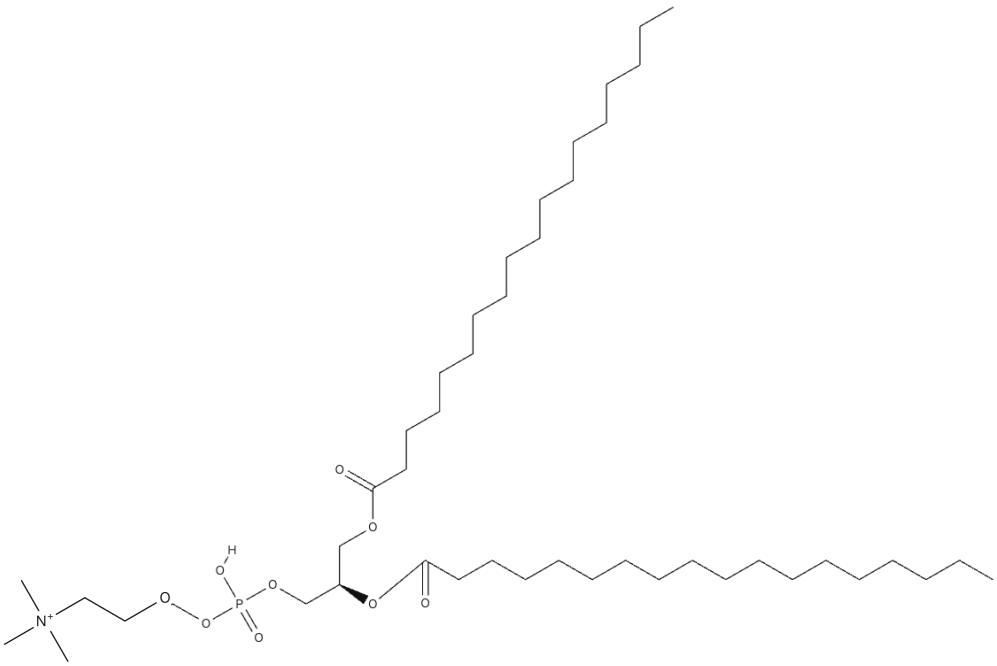


図 6: ホスファチジルコリン

ホスファチジルエタノールアミン (phosphatidylethanolamine)

ホスファチジルエタノールアミン (PE) は, 動植物に広く存在するリン脂質で, 動植物では PC に次いで 2 番目に多く存在している. 動物細胞の形質膜では, コリン含有リン脂質では細胞の外側を向いて存在しているのに対し, PE の大部分は内側（細胞質側）を向いていることが示されている. その化学構造を図 7 に示す.

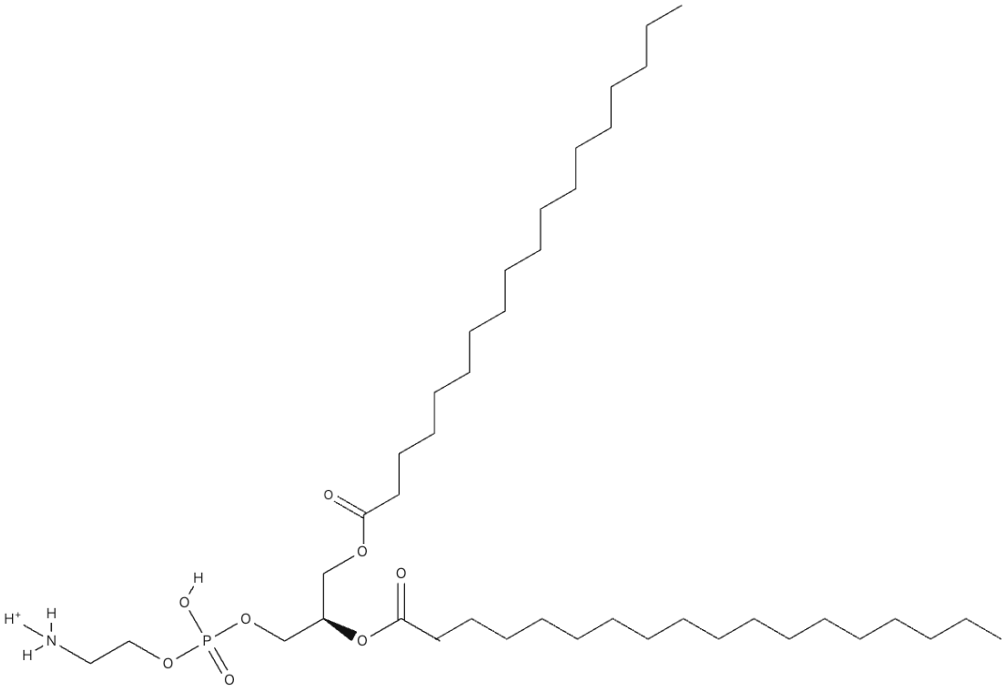


図 7: ホスファチジルエタノールアミン

ホスファチジルセリン (phosphatidylserine)

ホスファチジルセリン (PS) は, アミノ酸含有リン脂質として, 生体内に広く分布しているリン脂質の一つで, 特に脳や神経に多く存在している. その化学構造を図 8 に示す.

ホスファチジルイノシトール (phosphatidylinositol)

ホスファチジルイノシトール (PI) は, 広く動植物界に分布しているリン脂質の一つである. 植物細胞やカビでは主要なリン脂質成分であるが, 動物細胞では全リン脂質の数%に過ぎない. 動物細胞では, PI は形質膜の内層に多く分布しており, アラキドン酸の含量が高い. 細胞機能と PI の関連が研究より, 細胞外からの刺激に対し, PI をはじめとするイノシトールリン脂質は, 速やかに代謝され, 細胞内情報伝達機構において重要な役割を担っていることがわかっている.

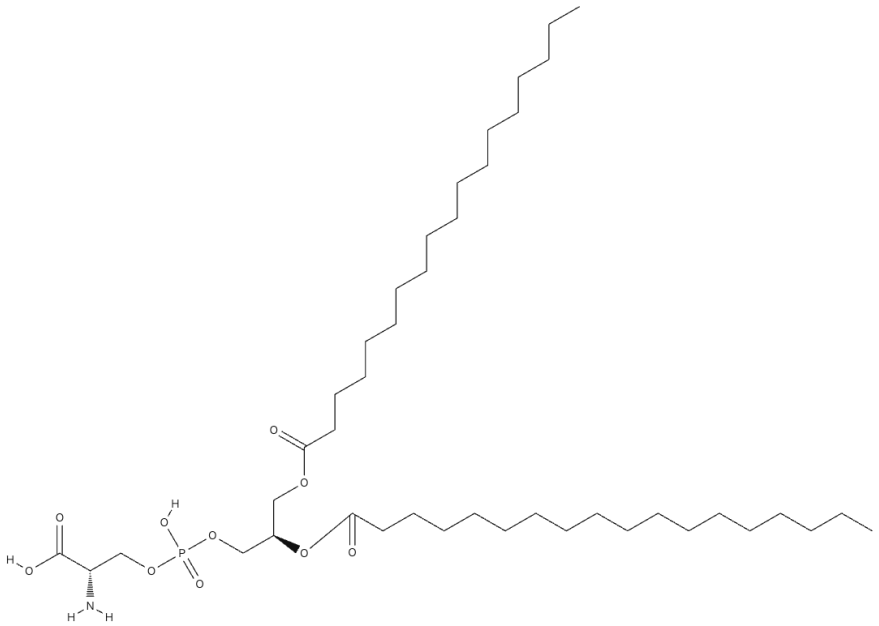


図 8: ホスファチジルセリン

イノシトールリン脂質は細胞機能発現に重要である. その化学構造を図 9 に示す.

ホスファチジン酸 (phosphatidic acid)

ホスファチジン酸 (PA) は, 生体膜脂質の 1 %以下にすぎない微量成分である. その化学構造を図 10 に示す.

ホスファチジルグリセロール (phosphatidylglycerol)

ホスファチジルグリセロール (PG) は, 植物から発見されたリン脂質であり, 現在では膜成分として生物界に広く分布することが知られている. その化学構造を図 11 に示す.

カルジオリピン (cardiolipin)(**ジホスファチジルグリセロール** (diphosphatidyl glycerol))

カルジオリピン (CL) は, 最近から高等動植物まで広い範囲の生物に含まれているリン脂質であり, 一般に不飽和脂肪酸の含有量が多い. その化学構造を図 12 に示す.

スフィンゴミエリン (sphingomyelin)

スフィンゴミエリンは, Thudichum により脳の白質に多量に発見され, スフィンゴシン, 脂肪酸, リン酸, コリンから構成されていることが 1884 年にわかった. 高等動物の細胞膜や神経組織ミエリンなどに広く分布しており, リン脂質としては, グリセロリン脂質と違い唯一のスフィンゴリン脂質である. その化学構造を図 13 に示す.

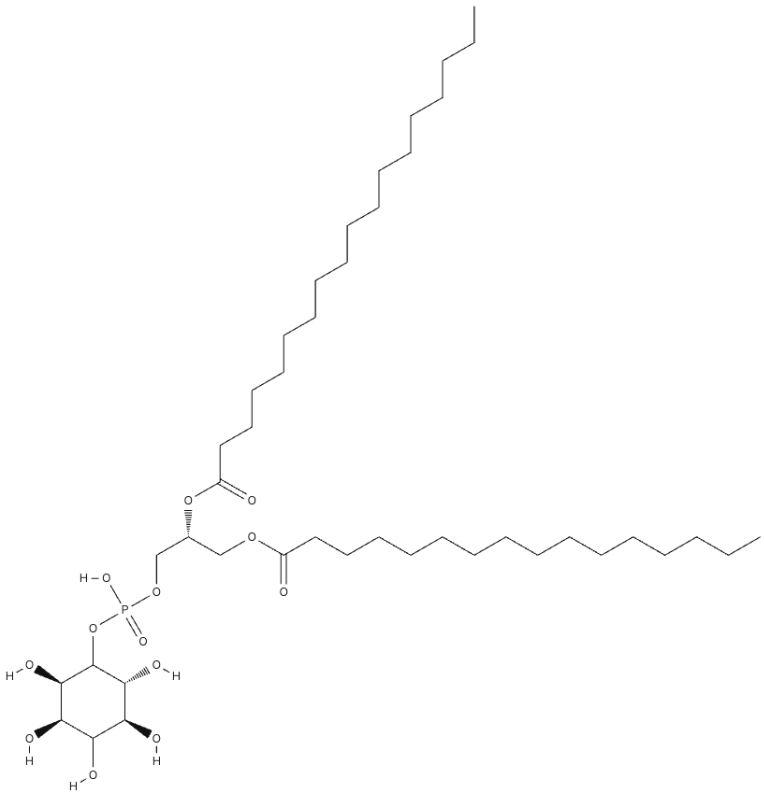


図 9: ホスファチジルイノシトール

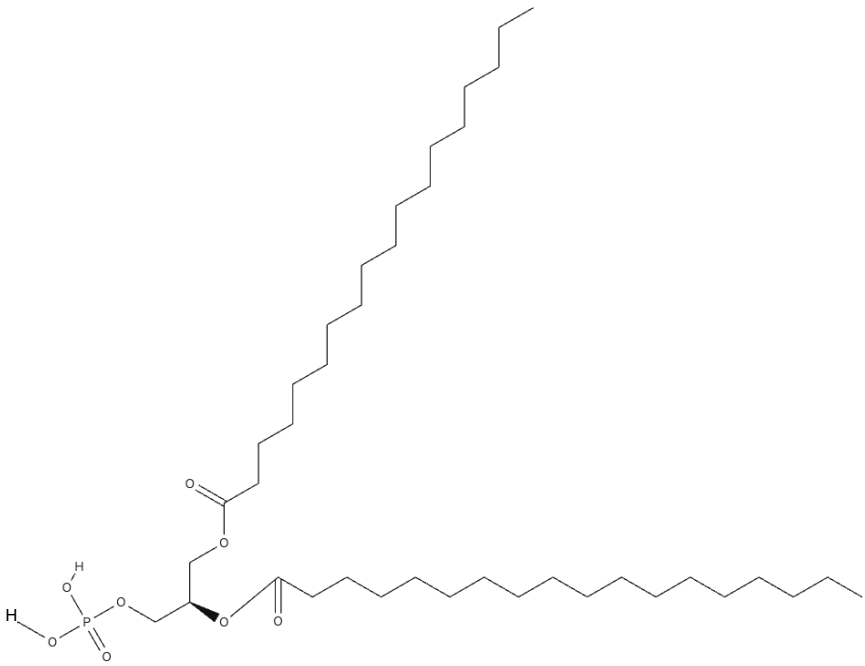


図 10: ホスファチジン酸

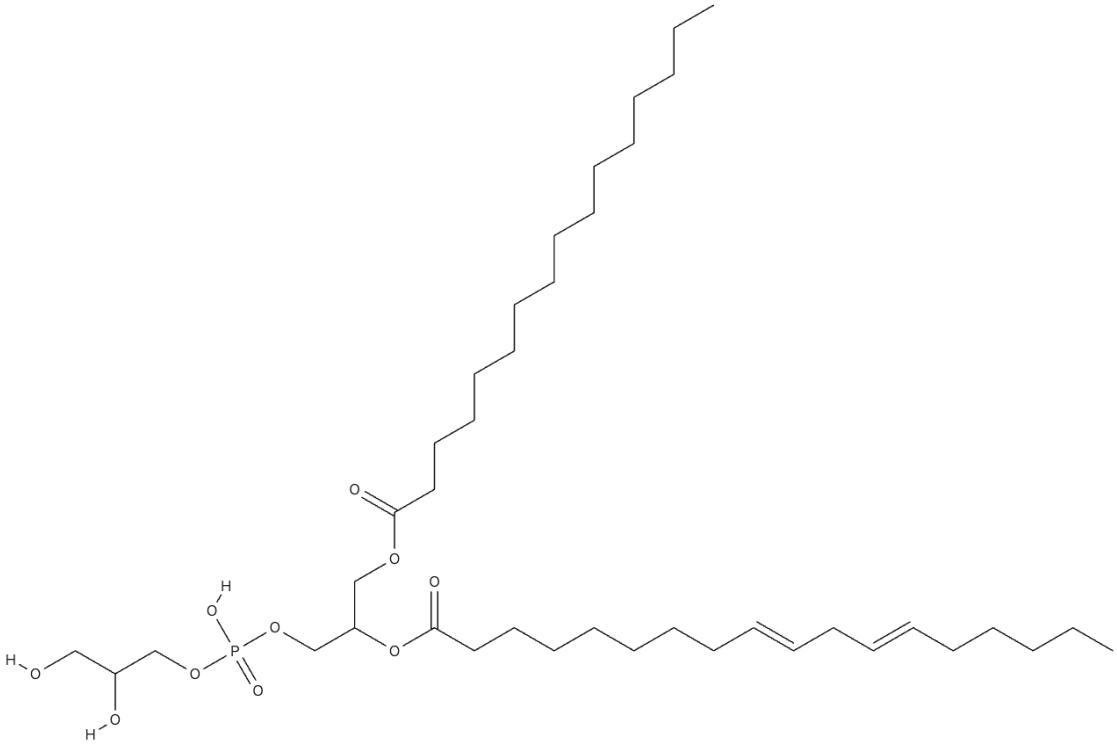


図 11: ホスファチジルグリセロール

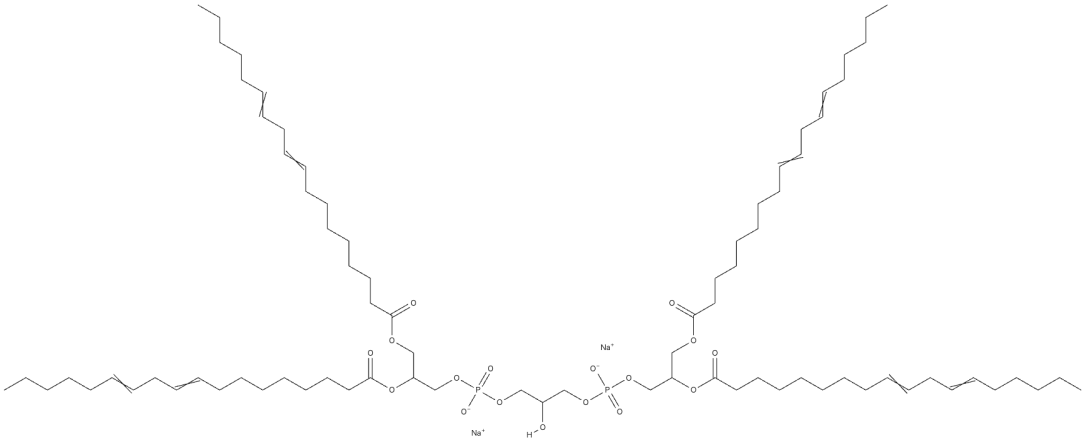


図 12: カルジオリピン

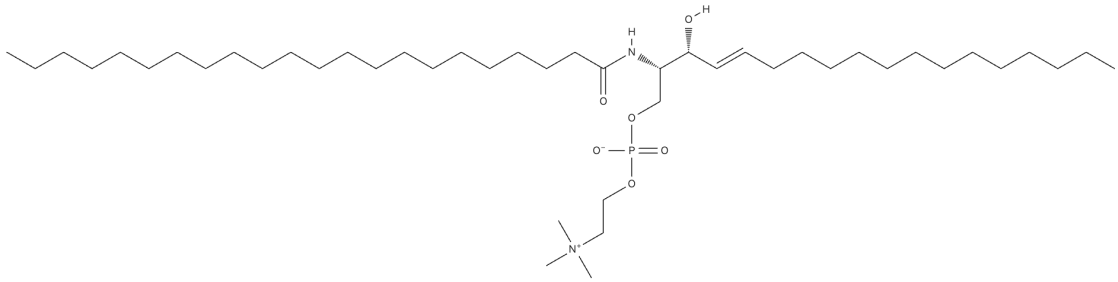


図 13: スフィンゴミエリン

3.2　生体内でのグリセロリン脂質生合成

生物の細胞を囲む生体膜は, 細胞機能を保つ上で非常に重要である. グリセロリン脂質は, 前節で述べたように細胞膜内外に広く分布し, 細胞の活性に重要な役割を果たしている. 生体膜グリセロリン脂質の代謝は厳密に制御されており, 常に生合成され続ける必要がある. 動物細胞における主要なリン脂質の代謝経路 [4] を図 14 に示す. 動物細胞における主要な膜リン脂質

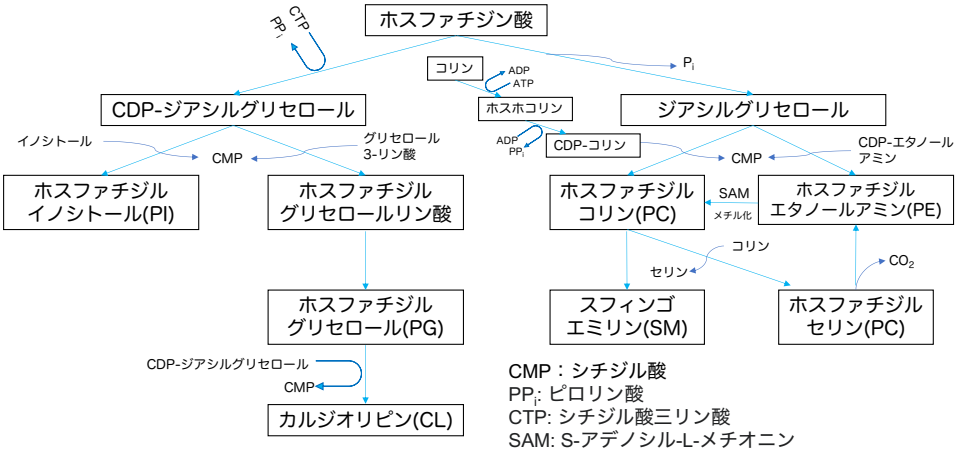


図 14: 動物細胞の主要リン脂質代謝経路

ホスファチジルコリン (PC) は, おもに CDP (Cytidine diphosphate, シチジン 2 リン酸, 図 15)-コリン (de-novo) 経路により生合成される. 肝以外の組織ではすべての PC が CDP-コリン経路により合成されると考えてよい. 1955-56 年, Kennedy と Weiss はラット肝膜画分を用いた実験で, ホスホコリン (phosphocholine, 図 17) からのホスファチジルコリンの合成に,

1. CTP (Cytidine triphosphate, シチジン 3 リン酸, 図 16) が必要である
2. 中間体として CDP-コリンを経由する
3. 1,2 - ジアシルグリセロールが直接の前駆体になる

ことを明らかにしていた. コリンホスホキナーゼ（コリンキナーゼ）は, コリンのリン酸化を触媒する酵素で, 種々の動物臓器に存在する事実がすでに Wittenberg と Kornberg [5] により報告されていた. これらから PC の主要合成経路＝ CDP-コリン経路の存在が明確になった. 同様の合成経路がホスファチジルエタノールアミン (PE) の生合成にも存在することを明らかにしており [6], これらの主要合成経路を発見者にちなんでケネディー経路 (Kennedy Pathway) と呼ぶ.

ケネディー経路は de novo 経路 (de novo: (ラテン語で) 初めから, 新たにの意味.) であり, 解糖系で得られるグリセロール-3-リン酸 (G3P, 図 18) から合成される. しかし, この経路では, すべ

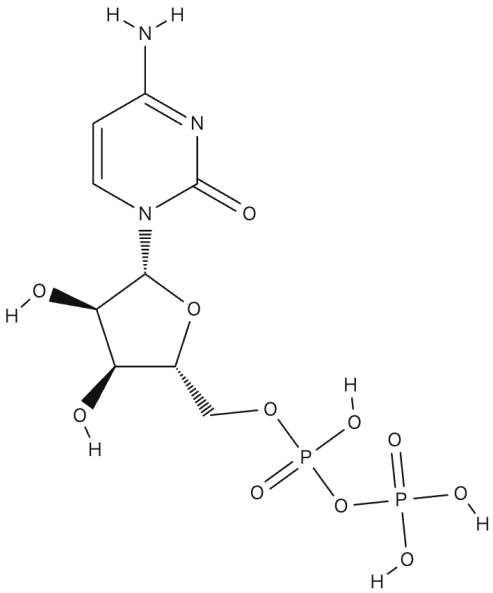


図 15: CDP

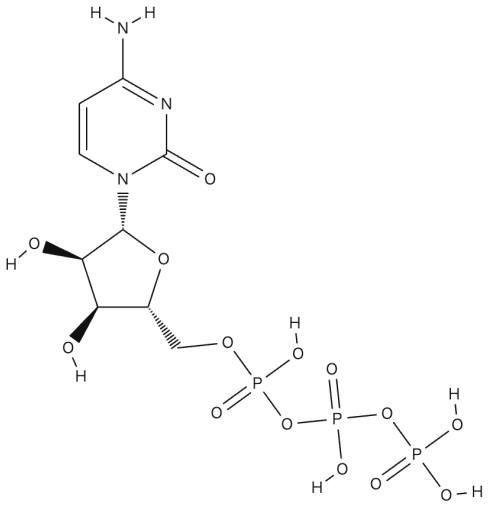


図 16: CTP

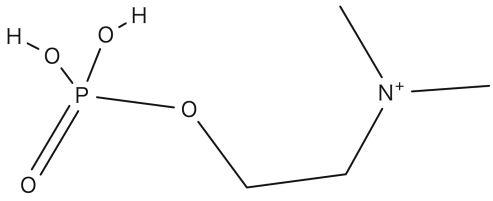


図 17: ホスホコリン

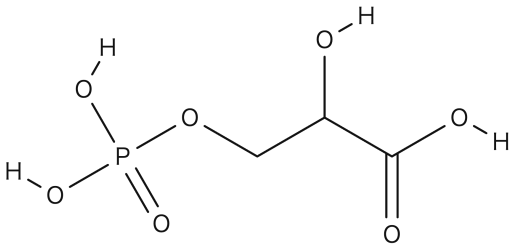


図 18: G3P

ての生体膜グリセロリン脂質の多様性を説明できない. そこで新たに提案されたのが, リモデリング経路であるランズ回路である [4]. ランズ回路は, 生体膜中のグリセロリン脂質を生合成する経路である (図 19). この回路では, ケネディー経路で合成されたグリセロリン脂質の sn-2 位 (グリセロール骨格の 2 位, 多価不飽和脂肪酸がエステル結合している) をホスホリパーゼ A2 によって切断することで, リゾリン脂質を作る. その後, アシル転移酵素 (AT) によって, 脂肪酸がリゾリン脂質に再結合されることで, sn-2 位にさまざまな脂肪酸が結合されたリン脂質が生合成される. このように, ランズ回路は複雑で多様な生体膜グリセロリン脂質の生合成を説明することができる.

ランズ回路の酵素は, ゲノムデータベースの充実により研究が進んでおり, 十数種類が現在までに報告されている. その 1 つである酵素 LPCAT4 [7] は, 特にマウスの脳, 精巣上体, 精巣, 卵巣に強く発現することが知られており (図 20), この酵素をノックアウトすることで, マウス

精巣中の精子形成に影響を及ぼすものと予想される.

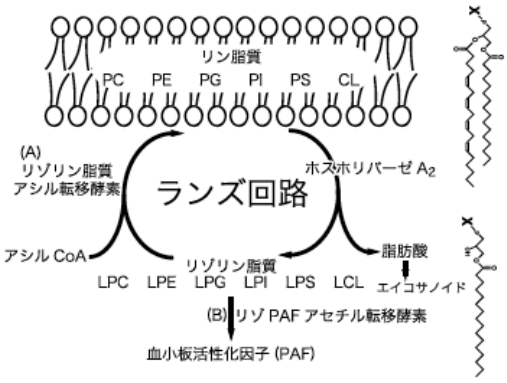


図 19: ランズ回路の概略. 生化学 第 82 巻 第 12 号, pp. 1091-1102, 2010

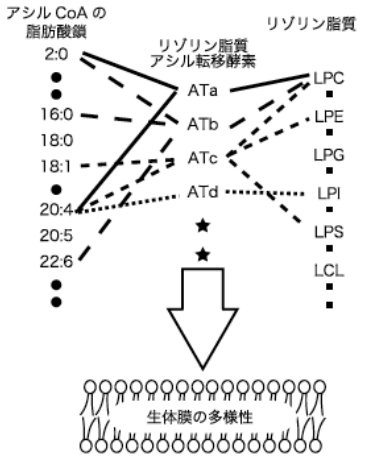


図 5 リゾリン脂質アシル転移酵素による生体膜多様性形成
一つのリゾリン脂質アシル転移酵素が，複数のアシル CoA と複数のリゾリン脂質を基質として様々なグリセロリン脂質を合成する．各組織に特徴的な生体膜多様性を生み出す．供給される基質の種類によっても変わるかもしれない．この図ではアシル転移酵素を ATa，ATb，★などと表記している．他のアシル CoA の脂肪酸鎖を●，他のリゾリン脂質（結合様式，脂肪酸種）を■と表記した．

図 20: LPCAT の生体膜中の働き. 生化学 第 82 巻 第 12 号, pp. 1091-1102, 2010

3.3 精巣の構造

マウス精巣全体の断面図を, 図 21 に示す. 精巣は, 精子形成を行う精細管と, ホルモン産生を行う間細胞 (ライディヒ細胞) からなる. 哺乳類の精巣内では, 少数の幹細胞が自己複製を繰り返すことで, 分化細胞である精子を生み出し続ける. マウスの精子形成は, 精細管 (seminiferous tubule) と呼ばれる長い管の内部で行われ, 特に, 曲精細管 (convoluted seminiferous tubule; CST) とよばれる, 迂曲し複雑に折り畳まれた精細管の中で発達する. 精原細胞とよばれる, 体細胞分裂をする段階の細胞は, 精細管内側の基底膜の上に存在する. 減数分裂を経て 1 次精母細胞, 2 次精母細胞, 精子細胞 (半数体) と, 分化が進行するにつれ, より中心の内腔へと移動する.（図 22, 23）

セルトリ細胞 (Sertoli cell)

セルトリ細胞は不安定の核を持つ, 精上皮の基底側から管腔に向かって伸びた極性を示す柱状で大型の体細胞で, 精細胞の分化に大きく関わっている. セルトリ細胞の役割としては, 精細胞の機械的な支持, 精細胞への栄養供与, 種々のタンパク質の分泌, 精子離脱の補助, 貪食作用,

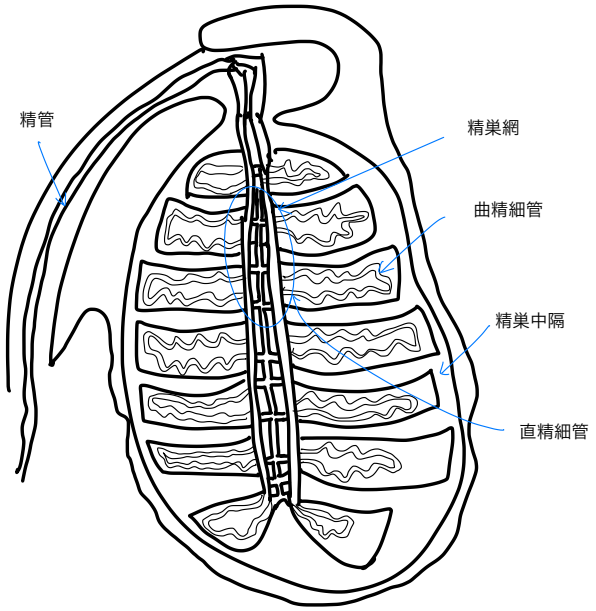


図 21: マウス精巣全体の断面図

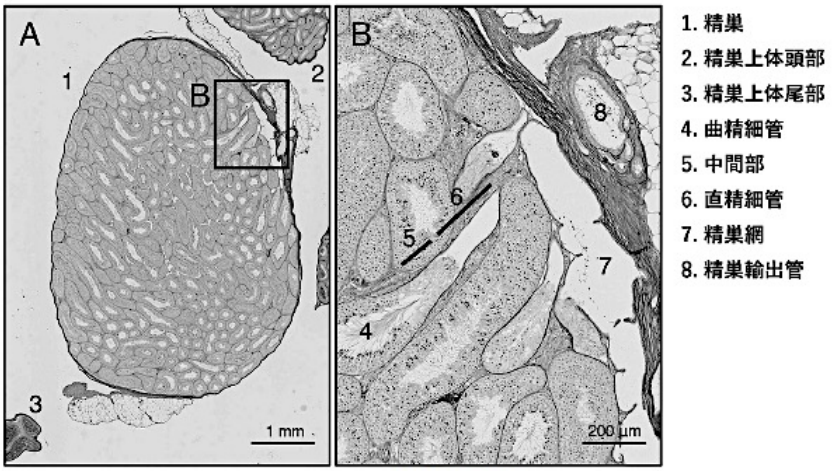


図1. マウス成体精巣の全体像.

図 22: マウス精巣の全体像.
金沢大学十全医学会雑誌　第 125 巻　第 3 号, 119-123(2016)

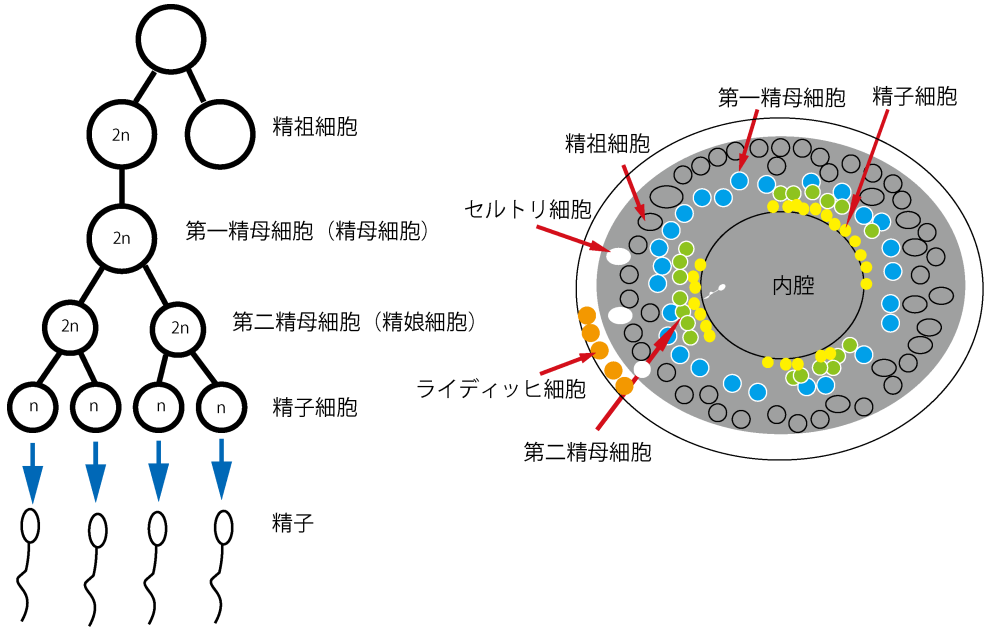


図 23: マウス精巣断面図と詳細

精細胞の免疫学的障壁などさまざまである. 各段階の精子形成細胞でセルトリ細胞と接触し, その影響を大きく受けている [8].

間質（内分泌）細胞 (Interstitial(endocrine) cell)（ライディッヒ細胞;Leydig's cell)

精細管の間を占める間質領域は, 間質細胞, マクロファージ, 線維芽細胞などによって構成されている. ライディッヒ細胞は不規則な形で様々な大きさを持っており, 周囲に突起を伸ばしているものもある.

3.4　精子形成の周期性

CST 内の精子には, 周期性があることが指摘されている. CST の断面から, ある CST 内には, 特定の発達段階の精子細胞のみが存在し, また他の CST 内には, 異なる段階の精子細胞のみが多く存在することがわかる. それぞれの CST 断面では特有の発達段階の細胞のみが見られ, それらは CST 内で周期的に現れる.

3.5　精子形成不全

ドコサヘキサエン酸 (docosahexaenoic acid; DHA) は, 脳機能や生殖機能に重要な役割を果たしている. DHA は血中でタンパク質と結合する形で存在するだけでなく, 生体膜中では DHA

含有リン脂質という形で取り込まれている. DHA 含有リン脂質生合成酵素 (LPCAT3) を欠損させたマウスでは, DHA 含有リン脂質の顕著な減少により, 視覚機能や雄性生殖機能が失われていることがこれまでの研究により判明している [9]. このような結果から, リン脂質に結合した DHA は視覚や生殖に必須であると言える.

4　MSI データに対する多変量解析と機械学習

広義に機械学習とは, コンピューターによりデータから規則や知識を抽出することを意味する. 本研究においては, 多次元の数値データからコンピュータープログラミングを用いて生物学的な知見を発見するために機械学習を用いる.

4.1　教師なし学習と教師あり学習

あらかじめラベル付けした（正解を与えた）データを訓練データとして用い, 入力から正解を出力するための関数を生成する手法を教師あり学習 (supervised learning) と呼ぶ. 出力が実数値である場合を回帰, 離散的なクラスである場合を分類と呼ぶ. 教師あり学習に対し, 入力データに対する正解が存在せず, 入力データに存在する未知の確率分布を, 何らかの形で学習する方法を教師なし学習 (unsupervised learning) と呼ぶ. この方法には, 多次元データから重要な情報を抽出する特徴抽出や次元削減, クラスタリングなどがある.

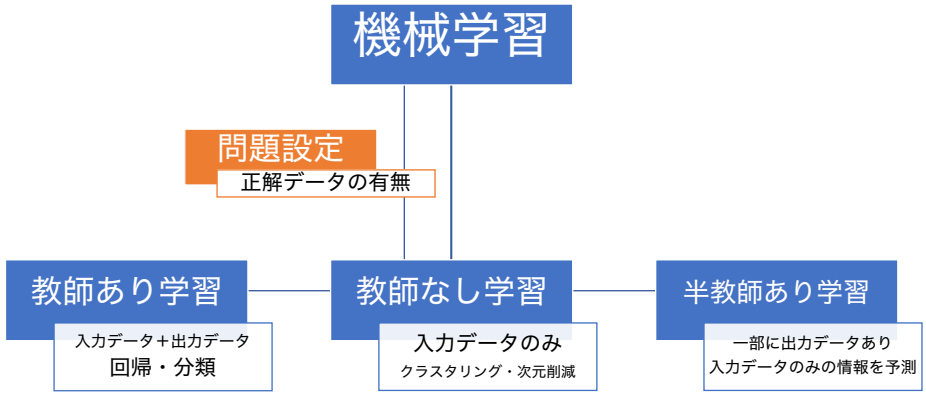


図 24: 機械学習の種類

表 1: 機械学習の手法例

手法	教師あり or なし	特徴
PCA （主成分分析）	なし	次元削減
k-means	なし	クラスタリング
t-SNE	なし	多次元データを圧縮, より高度にクラスタリング
線形判別分析	あり	次元削減, 識別
SVM	あり	超平面により識別, マージン最大化による汎化

4.2 主成分分析

高次元データを扱う際に, 各変数同士の相関を低次元で可視化することを考える. このような場合に, 次元を削減する方法として主成分分析 (principal component analysis; PCA) が用いられる. PCA とは, 元のデータの特徴を最もよく表現する新たな軸・主成分 (Principal Component; PC) を構成する方法である. つまり, 第 1 主成分は元データの分散が最も大きくなるようにとり, 第 2 主成分は第 1 主成分に直交するようにとることになる. このようにして主成分を元のデータの次元数と同じ数だけ構成することができ, 各主成分が全体のデータの分散に占める割合・寄与率 (contribution rate) をもとに, 主成分数を決定する. どのように主成分数を決定するかは明確な指標はなく, 各分析あるいは研究者によって異なるが, 今回は, 寄与率の合計・累積寄与率 (Cumulative contribution rate) を考慮し, 第 5 主成分まで採用することとした. [10]

PCA の具体的な手順は次のとおりである. 簡略化のため, データ行列 $\boldsymbol{X}$（$\boldsymbol{X}$ の各行はピクセル, 各列は m/z）中のデータは標準化 (Average = 1, Variance = 0) されていると考える. このとき, 共分散行列 Σ は

$$\Sigma = \frac{\boldsymbol{X}^T \boldsymbol{X}}{n-1} \tag{1}$$

と定義できる. ただし, $\boldsymbol{X}^T$は行列 $\boldsymbol{X}$ の転置行列を表す. PCA はこの共分散行列 Σ の固有値問題を解くことと一致する. Σ を

$$\Sigma = \boldsymbol{V}_p \Lambda \boldsymbol{V}_p^T \tag{2}$$

と分解すると ($\boldsymbol{V}_p$: 固有ベクトル, Λ : 対角行列), この固有ベクトル $\boldsymbol{V}_p$ の各列 $\boldsymbol{V}_i$ が, 各主成分（ローディングベクトル）であり, 対角行列 Λ の各対角成分 λ_i は, 各主成分の分散となる. 各主成分が元データ X をどれだけ含んでいるかを表す主成分スコアは, $\boldsymbol{V}_p^T X$ として計算できる. また, 特異値分解 (singular value decomposition; SVD) を用いて元のデータ X を行列分解すると

$$\boldsymbol{X} = U \Sigma \boldsymbol{V}_s^T \tag{3}$$

のように分解できる. ここで, $U, \boldsymbol{V}_s$ はユニタリ行列である. これを用いて $X^T X$ を計算すると

$$\begin{aligned} X^T X &= (U \Sigma \boldsymbol{V}_s^T)^T (U \Sigma \boldsymbol{V}_s^T) \\ &= \boldsymbol{V}_s \Sigma U^T U \Sigma \boldsymbol{V}_s^T \\ &= \boldsymbol{V} \Sigma^2 \boldsymbol{V}_s^T \end{aligned} \tag{4}$$

式 (2) と式 (4) を比較すると, 本質的に一致していることから, PCA は数学的には特異値分解と似た問題を解いていることとなる.

4.3 クラスタリング

クラスタリングとは, 教師なし学習として, 入力データをそれらの間の類似度に基づき, 複数のクラスターに分割する方法である. 行 (ピクセル, 座標) をグループとしてクラスタリングす

る場合, 類似のサンプル点を見つけることができる. 列 (m/z, 変数) をグループとしてクラスタリングする場合, 類似の脂質の発見や, 代表的な脂質の選択が期待される.

4.3.1 k-平均法

k-平均法 (k-means clustering) は代表的なクラスタリング法である. データ点を k 個のクラスタにランダムに割り当てることで初期化を行い, そのクラスタ平均を計算する. その計算値をもとにクラスタの割り当てを再度行う作業を, 繰り返し行いながら, 最適なクラスタリングを行うものである.(図 25, 26)

k-平均法の流れ

1. クラスタの数 k を決定する.
2. データ点をランダムに k 個のクラスタに分類する.
3. それぞれのクラスタ平均を計算する.
4. 各データ点と k 個のクラスタ平均との距離を計算し, 最も近いクラスターに各データ点を分配し直す.
5. 上記 1〜4 の作業を, 各データ点の分配が終了するまで行う.

k-平均法は, 勾配降下法などと同様, 初期値依存性が高いので, 複数回分析を行った場合, それぞれで分析結果が異なる可能性がある.

k-means クラスタリングの一例を図 25, 26 に示す. ここでは, 2 次元データを乱数で 200 個生成し (図 25), これらを 3 つのクラスタに分類した.(図 26)

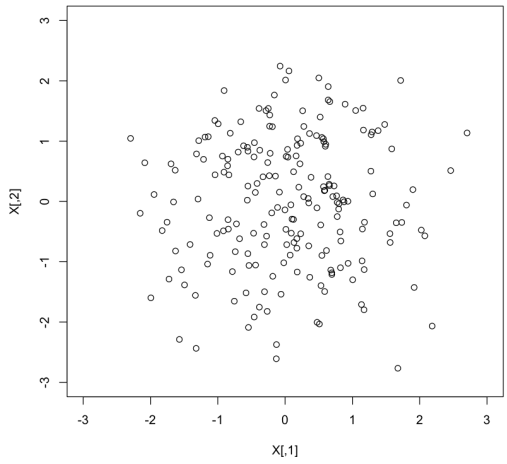


図 25: 元のデータ点

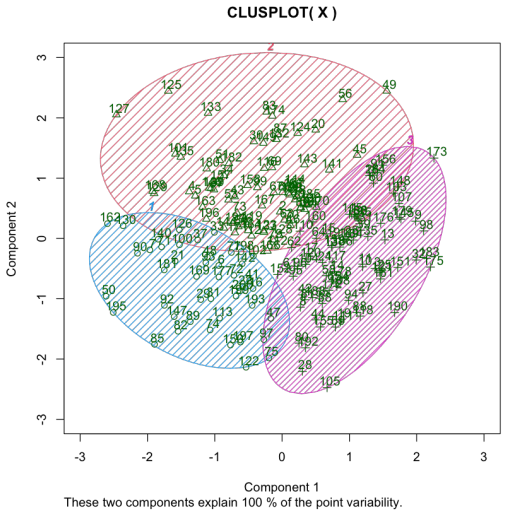
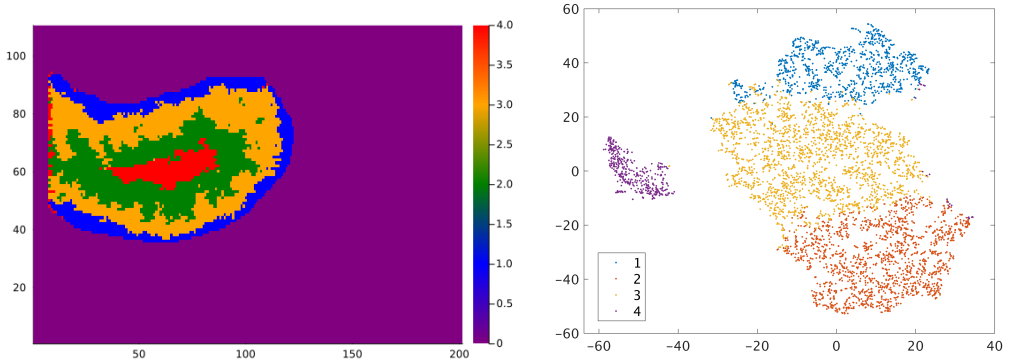


図 26: k-means clustering による分類 .

4.3.2　t 分布型確率的近傍埋め込み法

t 分布型確率的近傍埋め込み法 (t-distributed stochastic neighbor embedding; t-SNE) [11] は, 高次元データの可視化に適している次元削減アルゴリズムである. t-SNE では, データ間の類似度が反映されるように高次元のデータ点を低次元で表現することで次元を削減し, 高次元空間での近接するデータ同士を, 低次元空間でも近接点として表現できる. t-SNE によるデータクラスターの分離の例を図 27 に示す. ここでは, 7 章で用いたマウス精巣のデータ (野生型) を k-means クラスタリングによって 4 種類へと分離し, その分離度を検証すべく, ラベルをもとに t-SNE によるクラスタリングを同じデータに対し行なった.



(a) マウス精巣データの k-means clustering による 4 種類への分類 .

(b) (a) のラベルを元に, t-SNE によるクラスタリングを行なった結果. 4 種類がよく分離されていることがわかる.

図 27: t-SNE によるクラスタリングの例.

4.4　判別分析

前節で述べた PCA は, 入力データの次元を削減しデータの特徴を掴む, ‘教師なし学習 ‘に分類されるものであったのに対し, 判別分析 (discriminant analysis) は入力データと出力データがあらかじめ存在している, 教師あり学習に分類されるものである. 判別分析は, 入力データをいくつかのクラスへと識別するためのモデルを学習するものである. 本研究では, 出力データ（正解）として WT(正常), KO(異常) の 2 つを与え, 2 クラスへと分類することを目的とし, 判別分析の中で最も単純なモデルである, 線形判別分析 (linear discriminant analysis; LDA) を用いた.

LDA は, 最も 2 クラスをうまく識別できる直線で分類を行う方法である. 多次元データに対しては, 各データ点をある軸 $\boldsymbol{w}$ へと射影し, 射影した軸上で 2 クラス分類を行う. LDA では, この軸 $\boldsymbol{w}$ をデータから学習し作り出す. この軸 $\boldsymbol{w}$ は, 元の情報を射影したときに最もよく 2 クラスへと分類できるものであることから, LDA は次元削減の目的でも使用される.

射影軸 $\boldsymbol{w}$ の導出方法を次に示す [12]. まず最初に, 元データの各クラスの平均同士の差が最も大きくなるような決定方法を考える. 各クラスの平均値を $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ とおくと, 入力データ $\boldsymbol{x}$ の軸 $\boldsymbol{w}$ の射影を $y_n = \boldsymbol{w}^T \boldsymbol{x}_n$ と表せることより, 両クラスの差は

$$\boldsymbol{w}^T \boldsymbol{\mu}_1 - \boldsymbol{w}^T \boldsymbol{\mu}_2 \tag{5}$$

と定義できる. $||\boldsymbol{w}|| = 1$ という制約条件を加えると, ラグランジュ乗数 λ を導入することで, 評価関数 J を

$$J = \boldsymbol{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \lambda (\boldsymbol{w}^T \boldsymbol{w} - 1) \tag{6}$$

とおける. ラグランジュの未定乗数法により $\boldsymbol{w}$ で J を偏微分し, $J = 0$ とすると

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 + 2\lambda \boldsymbol{w} = \boldsymbol{0} \tag{7}$$

という条件が得られる. しかし, これだけでは不十分で, 各クラス内のデータの分布が極端に偏っている場合, 満足なクラス分けが不可能な場合が存在する. そこで, 分類後の各クラスの分散が最も小さくなるような条件をさらに考える. 各クラスの分散を V_1^2, V_2^2 とおくと, それぞれ

$$\begin{aligned} |\boldsymbol{V}_1|^2 &= \sum (\boldsymbol{w}^T \boldsymbol{x}_n - \boldsymbol{w}^T \boldsymbol{\mu}_1)^2 \\ |\boldsymbol{V}_2|^2 &= \sum (\boldsymbol{w}^T \boldsymbol{x}_n - \boldsymbol{w}^T \boldsymbol{\mu}_2)^2 \end{aligned} \tag{8}$$

となる. この合計が最も小さくすることが 2 つ目に考慮すべき条件である. ここで, クラス間変動行列 S_B, クラス間変動行列 S_W を

$$\begin{aligned} \boldsymbol{S}_B &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \\ \boldsymbol{S}_W &= \sum (\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \sum (\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T \end{aligned} \tag{9}$$

のように定義する. これらを用いて, 式 (6), 式 (8) の 2 つの条件を書き直し, 評価関数 J を再び考える. 各クラスの分散の和を最小にするという 2 つ目の条件は, J の分母に組み込むことができるから, J は

$$J = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{10}$$

のように書ける. この評価関数 J を最大化するような軸 $\boldsymbol{w}$ を求めれば良い.

4.5 サポートベクターマシン

線形判別分析に加え, 教師あり学習の手法として, サポートベクターマシーン (Support Vector Machine;SVM) を利用した. この手法は, 基本的に 2 クラス分類を目的として, 2 クラスを 1 つの超平面で分類する方法で, 超平面 $f(x) = wx + b$ が判別関数となる. SVM は, 超平面に最も近いデータ点, サポートベクター (Support Vector) によって構成される判別器と言える.

4.5.1 サポートベクターマシーンとは

サポートベクターマシーンとは, 学習データ $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_n}$ が 2 つに分けられていると仮定したとき, 変数 (w_1, w_2, ..., w_p) を法線ベクトルとする超平面 (Hyper Plane)

$$w_1x_1 + w_2x_2 + ... + w_px_p + b = \boldsymbol{w}^T\boldsymbol{x} + b = 0 \tag{11}$$

で全データを分類するという手法である. 線形分類可能なデータであれば, すべてのデータに対し

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b > 0 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b < 0 \end{aligned} \tag{12}$$

のいずれかに属すことになる. ここで

$$\begin{aligned} \boldsymbol{x}_1 \in G_1 \Longleftrightarrow \boldsymbol{w}^T\boldsymbol{x}_i + b > 0 \\ \boldsymbol{x}_2 \in G_2 \Longleftrightarrow \boldsymbol{w}^T\boldsymbol{x}_i + b < 0 \end{aligned} \tag{13}$$

として, G_1 に属するものを $y_i = 1$, G_2 に属するものを $y_i = -1$ とラベル付けすると, すべてのデータに対し

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0 \tag{14}$$

が成立する.

最もよくデータを 2 つに分類する方法とは, 線形分類可能なデータに対し, 最適な分離超平面を求めることである. 変数の係数ベクトル (重みベクトル) $\boldsymbol{w}$ と切片 (バイアス) b をどのように決定すべきかが問題となるが, ここでデータと超平面 H の間の距離 d を

$$d \equiv \frac{|w_1x_1 + w_2x_2 + ... + w_px_p + b|}{\sqrt{w_1^2 + w_2^2 + ... + w_p^2}} \tag{15}$$

と定める. 距離 d とは, 超平面 H から最も近いデータ点 (G_1に属するものを x_+, G_2に属するものを x_-とおく) までの距離である. この距離 d が最も大きくなるような超平面 H を設定することで, 最もよくデータを 2 つに分類することができる. この距離 d をマージンと呼ぶ. また, d は

$$d = \frac{\boldsymbol{w}^T\boldsymbol{x}_+ + b}{||\boldsymbol{w}||} = \frac{-(\boldsymbol{w}^T\boldsymbol{x} + b)}{||\boldsymbol{w}||} \tag{16}$$

と表せる. ここまでの議論から,

> 2 つのラベルを最もうまく分類する問題 $\Longrightarrow$ マージンを最大とする超平面を求める問題

と置き換えられる. マージン最大化の条件とは, すべてのデータ点の, 超平面 H からの距離が少なくとも d 以上であるような d を最大化すること, すなわち

$$\max_{w,b} d \; s.t. \text{ 条件 } \frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}{||\boldsymbol{w}||} \geqq d \; (i = 1, 2, ..., \; n) \tag{17}$$

となる. 一般に, この問題を直接解くのは計算上容易ではない.

4.5.2 2 次計画問題

ここで, 式 (17) の両辺を d で割ると

$$y_i(\frac{1}{d||w||}\boldsymbol{w}^T\boldsymbol{x}_i + \frac{1}{d||\boldsymbol{w}||}b) \geqq 1,\ i = 1, 2, ..., n \tag{18}$$

となり, $r = \frac{1}{d||\boldsymbol{w}||}$ として, 係数のスケールを変えると

$$y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + b^*)\ \geqq 1,\ i = 1, 2, ..., n \tag{19}$$

ただし, $w^* = rw,\ b^* = rb, x_+,\ x_-$ のデータについては, $y_i(\boldsymbol{w}^*\boldsymbol{x}_i + b^*) = 1$ より

$$d = \frac{y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + b^*)}{||w^*||} = \frac{1}{||\boldsymbol{w}^*||} \tag{20}$$

係数のスケールを変化させても, 超平面は不変であることから, スケール変換により, マージン最大化問題は

条件 $y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + b^*) \geqq 1$ のもとで $d^* = \frac{1}{||w^*||}$が最大となるような $\boldsymbol{w}^*$と b^*を求める

問題へと帰着できる. $1/||\boldsymbol{w}||$ の最大化を$\frac{1}{2}||\boldsymbol{w}||^2$ の最小化は等価なので, 主問題は

$$\min_w \frac{1}{2}||\boldsymbol{w}||^2\ s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqq 1\ \ (i = 1, 2, ..., n)$$

となる. これを 2 次計画問題という. サポートベクターマシーンでは, Lagrange 関数を導入し, 2 次計画問題を双対問題と呼ばれる最適化問題に帰着させて解く.

4.5.3 双対問題

制約条件の個数 n に相当する Lagrange 乗数 $\alpha_1,\ \alpha_2,\ ...,\ \alpha_n$を用い,

$$L(\boldsymbol{w},\ b,\ \alpha_1,\ \alpha_2,\ ...,\ \alpha_n) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n}\alpha_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)\} \tag{21}$$

を定める. 次に, L を $\boldsymbol{w}$, b で偏微分したものを 0 とおくと

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i = \boldsymbol{0} \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^{n}\alpha_i y_i = 0\end{aligned} \tag{22}$$

より

$$\begin{aligned}\boldsymbol{w} &= \sum_{I=1}^{n}\alpha_i y_i \boldsymbol{x}_i \\ \sum_{i=1}^{n}\alpha_i y_i &= 0\end{aligned} \tag{23}$$

が得られる. これらを元の式に戻すと

$$\begin{aligned} L_D(\alpha_1,\ \alpha_2,\ ...,\ \alpha_n) &= \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i + \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \end{aligned} \tag{24}$$

よって, 双対問題 (Lagrange 関数に関する最適化問題) に帰着させることができ,

$$\begin{aligned} \max_{\alpha_1,\ ...,\ \alpha_n} L_D(\alpha_1,\ \alpha_2,\ ...,\ \alpha_n) = \max_{\alpha_1,\ ...,\ \alpha_n} \{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \} \\ s.t.\ \alpha_i \geqq 0,\ \sum_{i=1}^{n} \alpha_i y_i = 0\ \ (i = 1,\ , 2,\ ...,\ n) \end{aligned} \tag{25}$$

この解を $\hat{\alpha}_1,\ \hat{\alpha}_2,\ ...,\ \hat{\alpha}_n$ とすると,

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i \tag{26}$$

という最適解を得る.
また, $x_+,\ x_-$ について

$$\begin{aligned} \hat{\boldsymbol{w}}^T \boldsymbol{x}_+ + \hat{b} = 1 \\ \hat{\boldsymbol{w}}^T \boldsymbol{x}_- + \hat{b} = -1 \end{aligned} \tag{27}$$

が成立するから,

$$\hat{b} = -\frac{1}{2} (\boldsymbol{w}^T \boldsymbol{x}_+ + \boldsymbol{w}^T \boldsymbol{x}_-) \tag{28}$$

よって, マージンを最大化する超平面 H は

$$H:\ \hat{\boldsymbol{w}}^T \boldsymbol{x} + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \hat{b} = 0 \tag{29}$$

この識別器による判別は, 未知のデータ $\boldsymbol{x}$ を用いた際に

$$\begin{aligned} \hat{\boldsymbol{w}}^T \boldsymbol{x} + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \hat{b}\ \geqq 0\ \rightarrow\ G_1 \\ \hat{\boldsymbol{w}}^T \boldsymbol{x} + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \hat{b}\ < 0\ \rightarrow\ G_2 \end{aligned} \tag{30}$$

のいずれに属するかによって分類できる.

4.5.4　線形判別分析と SVM の違い

線形判別関数の係数ベクトルは, 2 群すべての学習データから求めた標本平均ベクトルと標本分散共分散行列に基づいて推定される. これに対し, サポートベクターマシーンでは, 実質的に判別関数を構成するデータは最適な分離超平面を挟む 2 つの超平面上のデータ x_+, x_-, すなわち

$$y_i(\hat{\boldsymbol{w}}^T \boldsymbol{x}_i + \hat{b}) = 1 \tag{31}$$

を満たすデータのみである. このようなデータをサポートベクター (Support Vector) と呼び, サポートベクターによって構成される識別器が SVM であると言える . [13] SVM による識別の例を図 28, 29 に示す. サンプルは, フィッシャーのアヤメのデータを使用した. 図 28 には, ‘setosa‘と ‘virginica‘という 2 種類の花のデータを表示している. これを超平面により判別した結果が図 29 である. 図中で丸をつけたデータ点がサポートベクターであり, このデータ点によって超平面が構成されている.

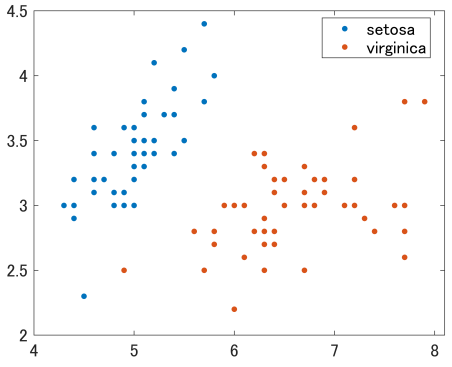


図 28: 2 種類のデータ点を含んだサンプルデータ

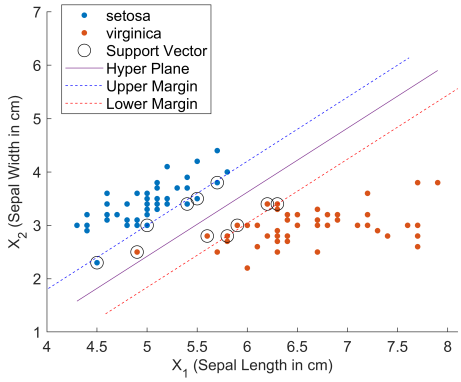


図 29: SVM の例. 超平面によって 2 種類のデータ点を識別している.

5　マウス精巣組織の計測

5.1　実験に使用したマウス

質量分析イメージングを行う生体試料として, 野生型 (Wild Type, 以降 WT) マウス精巣組織および酵素 LPCAT4 ノックアウト型 (Knock-Out, 以降 KO) マウス精巣組織を使用した. ここでは, まず正イオンモードで計測を行った. WT, KO それぞれの組織について, 1 ピクセルあたり 5μm × 5μm, 縦 1000μm, 横 550μm を 1 領域として (図 30), 5 領域測定した. 溶媒として, DMF と MeOH の 1:1 混合溶媒を用いた. 測定条件は以下の表 2 の通りである.

表 2: マウス精巣組織の測定条件.

測定試料・領域	振動周波数 [Hz]	電圧 [kV]	溶媒流量 [nL/min]
WT-1	721.660	4.00	5
WT-2	721.760	4.00	5
WT-3	721.712	4.00	5
WT-4	721.792	4.00	5
WT-5	721.750	4.00	5
KO-1	667.166	3.20	3
KO-2	667.442	3.30	3
KO-3	667.332	3.40	3
KO-4	667.454	3.40	3
KO-5	667.533	3.70	3

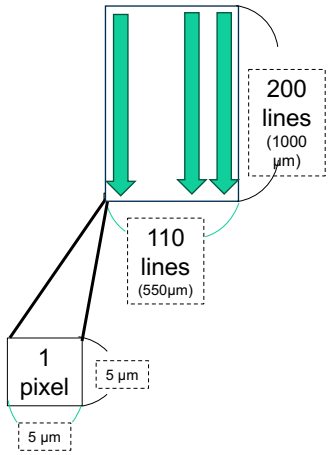


図 30: 測定のイメージ図. 試料ステージを移動させることで, 緑色の矢印方向にスキャンを行なった.

6　取得したデータの分析：WT 内, KO 内の脂質分布

前章で述べた, 計測したデータについて, 各分析を行った. 取得したデータは, 解析ソフト IMAGEREVEAL（島津製作所）を用い, 前処理を行った. 前処理として, モノアイソトピック質量のイオンピークの m/z の抽出, CST の領域のみに ROI(region of interest, 関心領域) を設定, ROI の全ピクセルのデータ行列の出力を行った. 自ら取得したデータである, 9/26 に取得した野生型 WT のデータと 10/5 に取得したノックアウト型 KO のデータをそれぞれ分析を行った. WT 内であっても, 各 CST 内における脂質分布が異なっていることは, 3.6 節で述べたとおりであるが, クラスタリングを行うことでその違いを導くことを目的とした. データ行列 (9/26, WT-POS1) は, それぞれ m/z 2998 次元, ピクセル数 21909 である.

6.1　k-means 法によるクラスタリング

まず, 9/26 に取得した野生型 WT のデータ (以降, 0926WT-POS と表す) について, k-means によるクラスタリングを行った. k-means 法は, 初期値によってクラスタリングの結果が異なる場合があるため, 複数回同様の操作を行った. 最初に, 3 種類にクラスタリングするように分析を行い. その結果を図 31, 32, 33 に示す. この結果から, 図中のラベルが 3(濃い緑色) の部分は CST 内部 (精子形成の最終段階), ラベルが 1（水色）の部分は精子形成の途中段階, ラベルが 2(黄土色) の部分は CST 外側 (精子形成の初期段階) にあることがわかる. よって, CST 断面の内部と外部で脂質の強度に明確な違いがあり, 図 31〜図 33 のそれぞれで分布が異なっていることから, CST 内の断面を切り取る場所によっても脂質強度が変化することが考えられる.

次に, 4 種類へのクラスタリングを行なった. その結果を図 34〜 36 に示す. 4 種類でのクラスタリングでは, より高精度な分類が期待されたものの, クラスター同士で重なり合う箇所が多数あった.

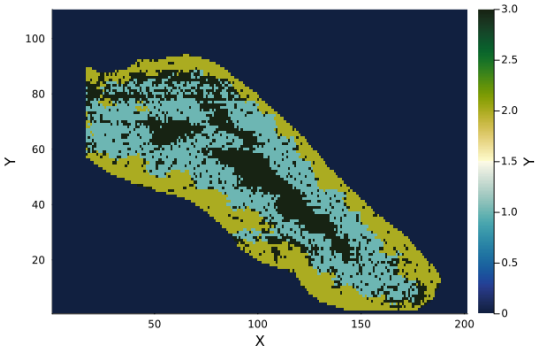


図 31: 0926WT-POS-CST1

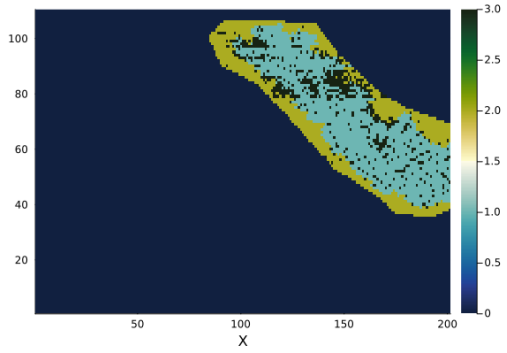


図 32: 0926WT-POS-CST2

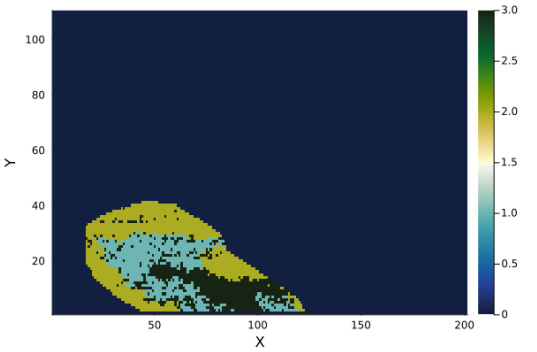


図 33: 0926WT-POS-CST3

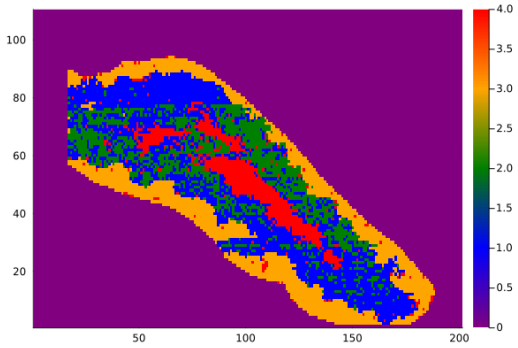


図 34: 0926WT-POS-CST1 について, 4 種類へのクラスタリング.

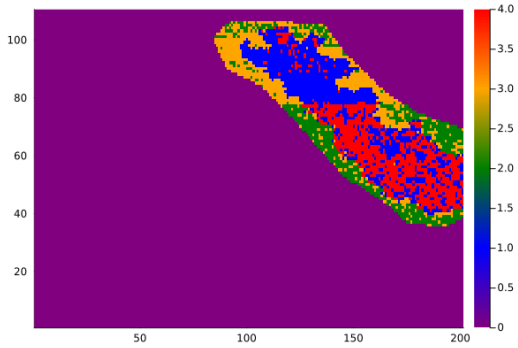


図 35: 0926WT-POS-CST2 について, 4 種類へのクラスタリング.

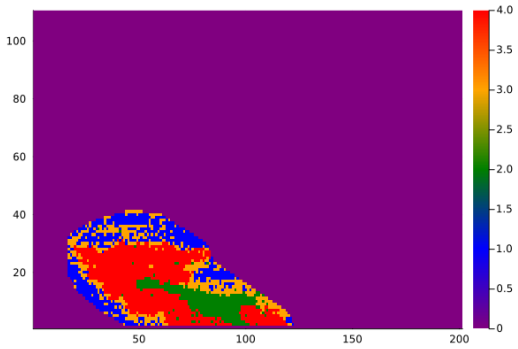


図 36: 00926WT-POS-CST3 について, 4 種類へのクラスタリング.

6.2　t-SNE によるクラスタリング

前節で kmeans クラスタリングを行なったが, そのクラスタリング精度を評価する目的で, t-SNE によるクラスタリングを行なった. ラベルは kmeans クラスタリングによって得たラベルを使用し, 3 種類でクラスタリングした場合, 4 種類でクラスタリングした場合の双方について, 0926WT-POS-CST1〜3 の t-SNE による分類を行った. その結果を図 37〜 39, 図 40〜 42 に示す. これらの結果から, 今回使用したデータに関しては 3 種類へのクラスタリングが最も適していると判断した.

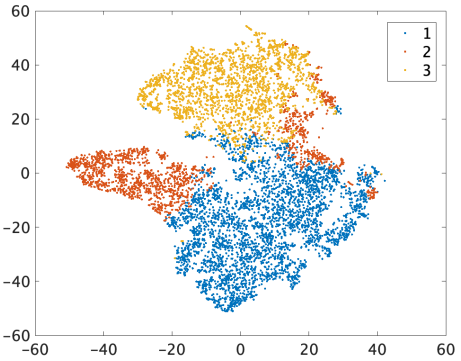


図 37: 0926WT-POS-CST1 の 3 種類へのクラスタリングについて, t-SNE を用いた結果.

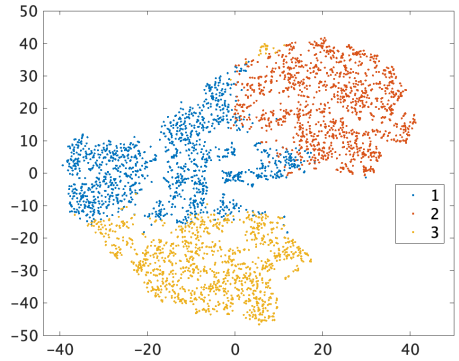


図 38: 0926WT-POS-CST2 の 3 種類へのクラスタリングについて, t-SNE を用いた結果.

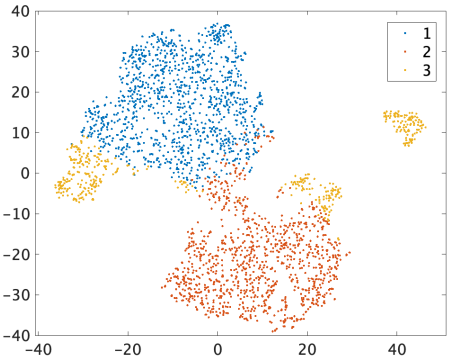


図 39: 0926WT-POS-CST3 の 3 種類へのクラスタリングについて, t-SNE を用いた結果.

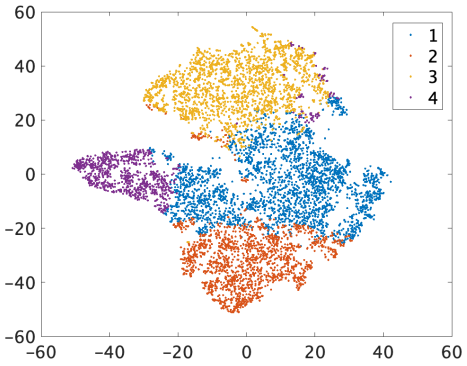


図 40: 0926WT-POS-CST1 の 4 種類へのクラスタリングについて, t-SNE を用いた結果.

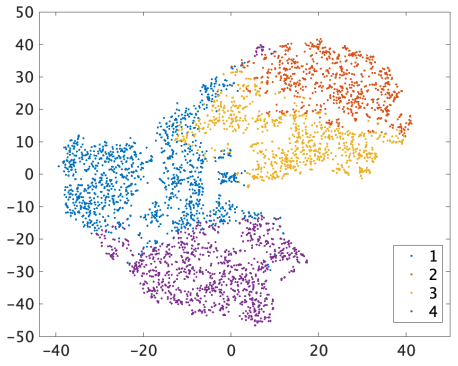


図 41: 0926WT-POS-CST2 の 4 種類へのクラスタリングについて, t-SNE を用いた結果.

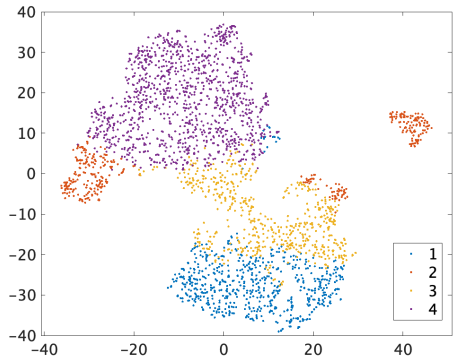


図 42: 0926WT-POS-CST3 の 4 種類へのクラスタリングについて, t-SNE を用いた結果.

7　計測結果の解析：WT と KO の判別

当研究室で取得されたデータを用いて, 解析を行った. このデータで用いたマウス精巣組織は, 野生型 (WT) と酵素 LPCAT3 ノックアウト型 (KO) を用い, 正イオンモードおよび負イオンモードで取得した.

7.1　PCA による分析

WT, KO の全データに対し PCA を適用した. 第 1〜9 主成分の寄与率を図 43 に示す. 図 43 中の折れ線グラフは, 累積寄与率を示しており, 今回は PC5 までを採用した. また, PC1〜5 の主成分スコアを縦軸, 横軸にとった散布図行列を図 44 に示す. PCA によって得られた各主成

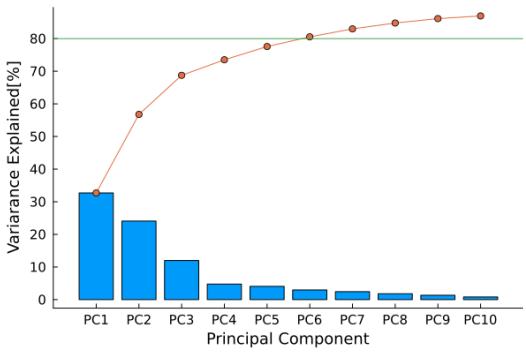


図 43: 第 1〜第 9 主成分の寄与率. 寄与率が 80 %に達したラインに緑線を引いている.

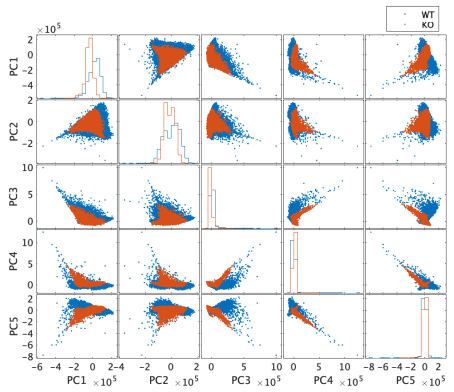
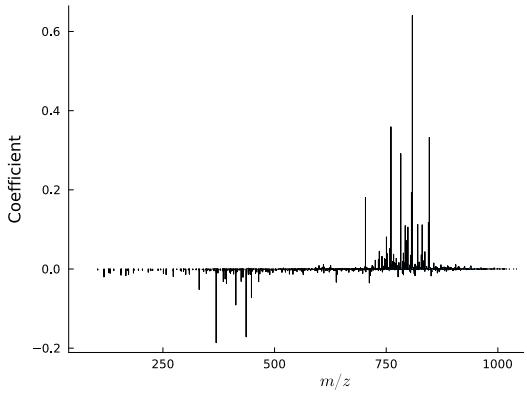
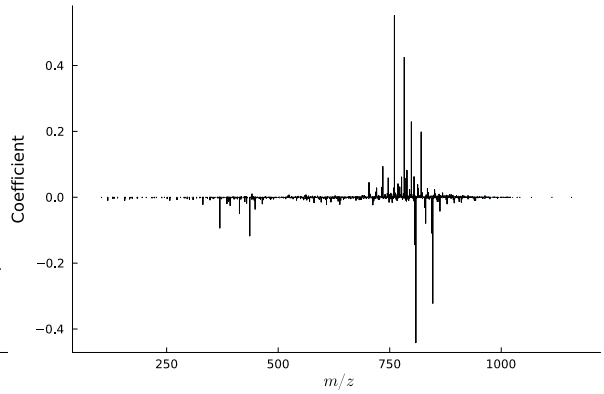


図 44: PCA による解析. PC1-5 についての散布図行列.

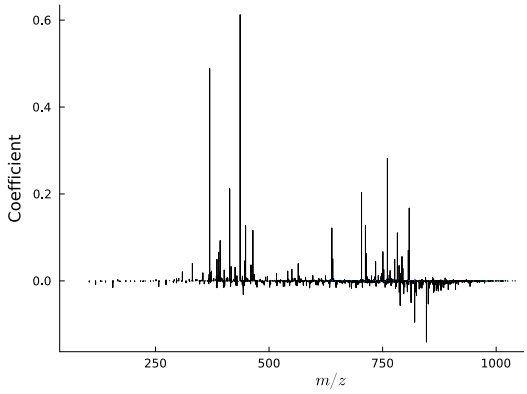
分の, 主成分係数（負荷量）を図 45a に示す. この数値は, 構成した主成分が, 各 m/z の元の情報量をどれだけ含んでいるかを示している. また, PCA によって得られたスコアを, 元の座標情報に反映させた PCA スコアイメージングを図 46a, 46b に示す. 主成分 1(PC1) と主成分 2(PC2) のスコアイメージングを見ると, CST の形が明確に示されており, CST 内と外の領域で脂質分布が大きく異なることがわかる. また, PC1 のイメージングから, CST の中心部分がスコアが小さくなっていることがわかるが, この部分は精子形成が完了し, 輸送される領域であることから, 脂質の強度が乏しいことが予想される. また, PC2 のイメージングでは, WT(野生型) では CST 内で階層構造のように脂質分布が異なっていることが確認できる, このことは, CST 内で外部から内部へと精子形成が成熟していくという事実を示していると考えられる. 一方, KO（ノックアウト型）では CST 内では一様に強度が低く, 均一なコントラストを示したことから, 酵素 LPCAT3 をノックアウトしたことにより, 計測した脂質において, 強度が低下し, 精子形成に何らかの問題が生じていると考えられる. また, PC3 以降のイメージングについて



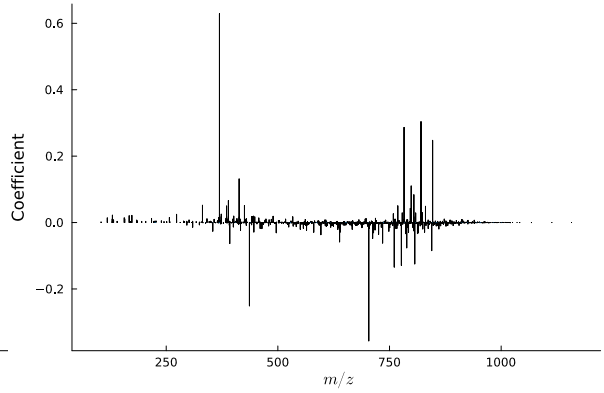
(a) PC1 の主成分係数.



(b) PC2 の主成分係数



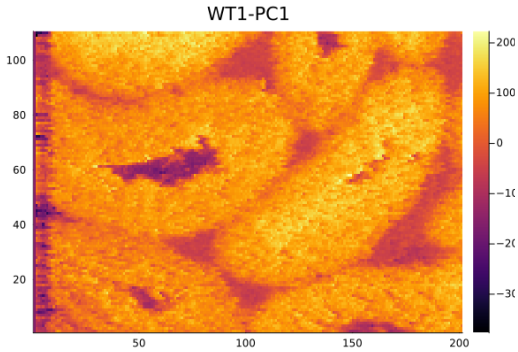
(c) PC3 の主成分係数



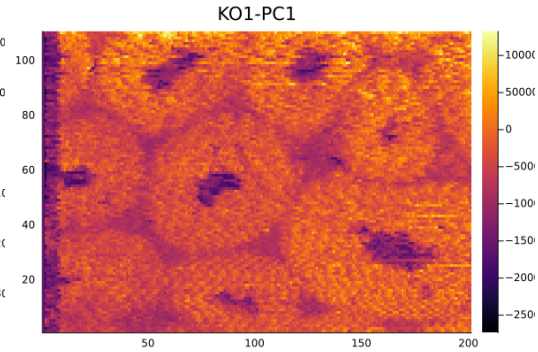
(d) PC4 の主成分係数

図 45: PCA により得た主成分係数の比較

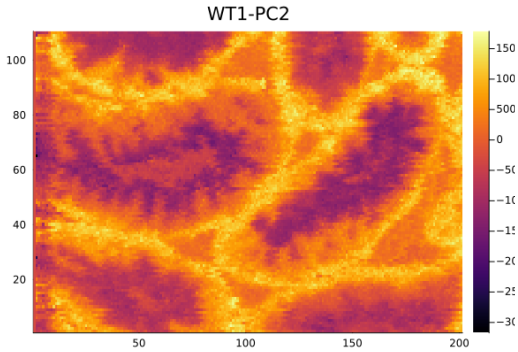
は, 全領域で強度の強弱が少なく, 得られる情報は少ない.



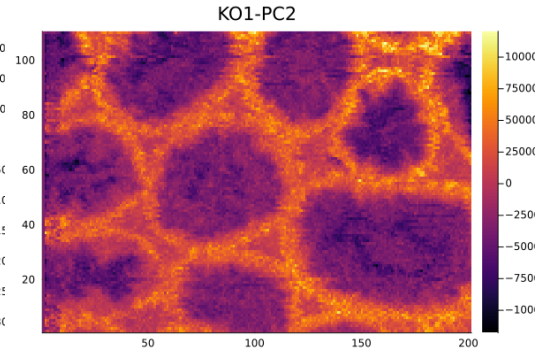
(a) PCA スコアイメージング-WT-PC1



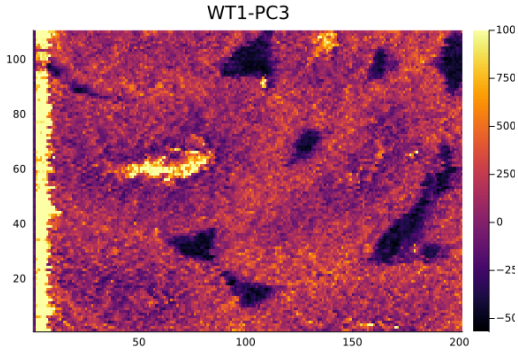
(b) PCA スコアイメージング-KO-PC1



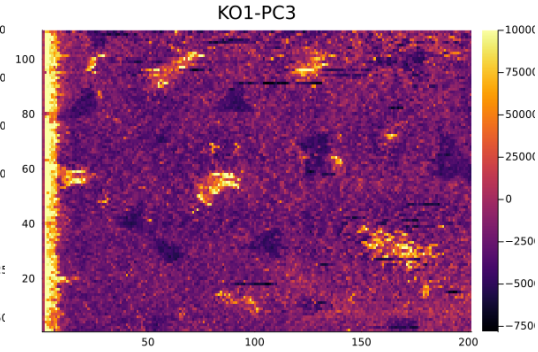
(c) PCA スコアイメージング-WT-PC2



(d) PCA スコアイメージング-KO-PC2

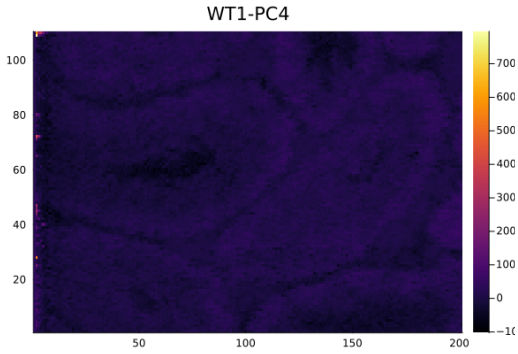


(e) PCA スコアイメージング-WT-PC3

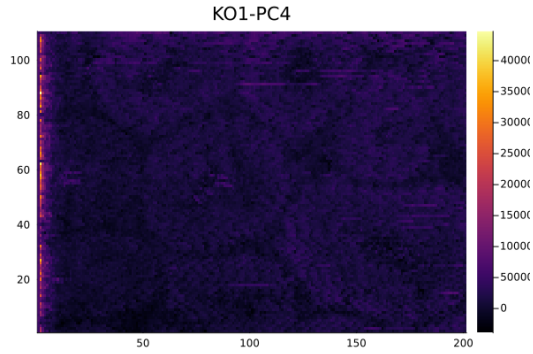


(f) PCA スコアイメージング-KO-PC3

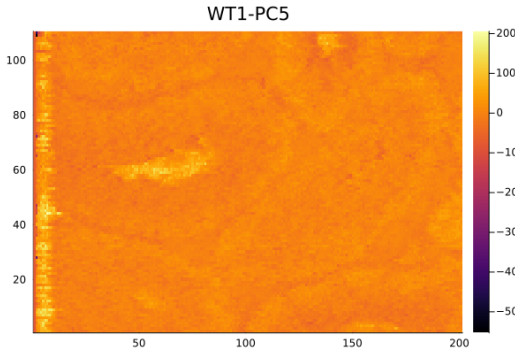
図 46: PCA スコアイメージの比較 (WT1, PC1-PC3)



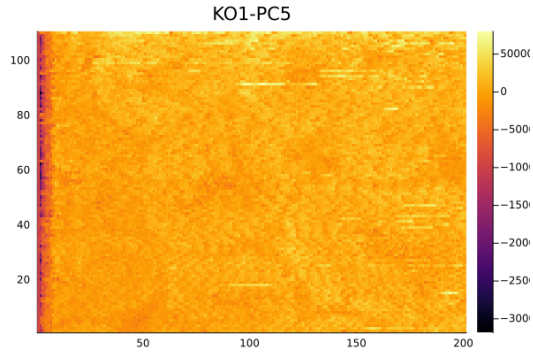
(a) PCA スコアイメージング-WT-PC4



(b) PCA スコアイメージング-KO-PC4

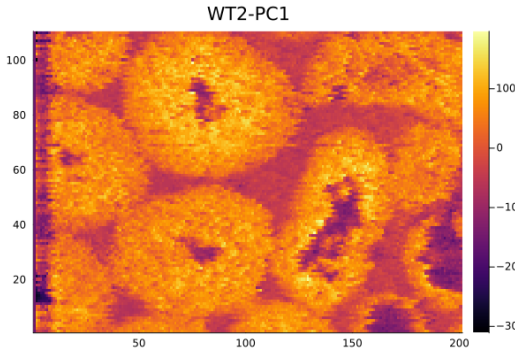


(c) PCA スコアイメージング-WT-PC5

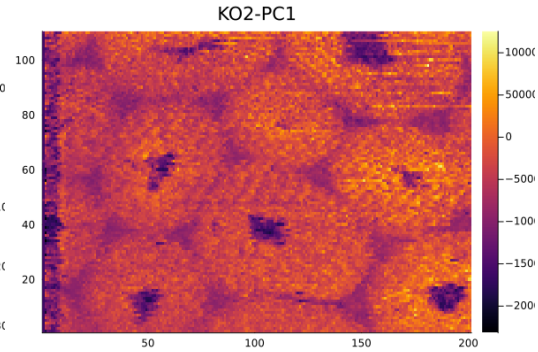


(d) PCA スコアイメージング-KO-PC5

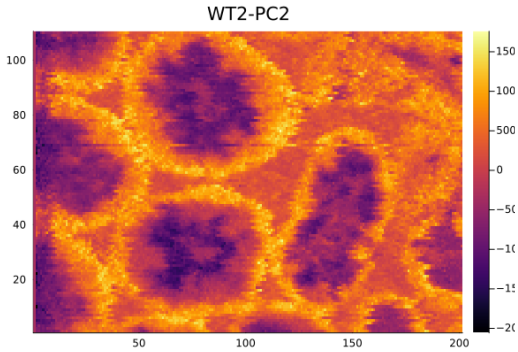
図 47: PCA スコアイメージの比較 (WT1, PC4-PC5)



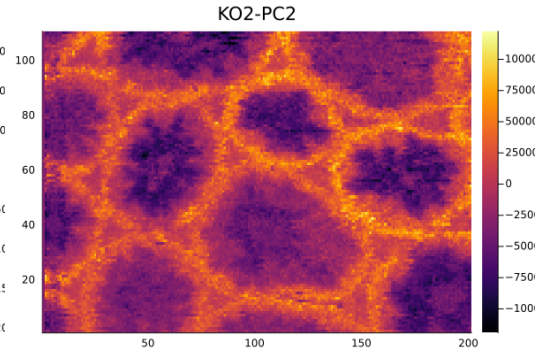
(a) PCA スコアイメージング-WT2-PC2



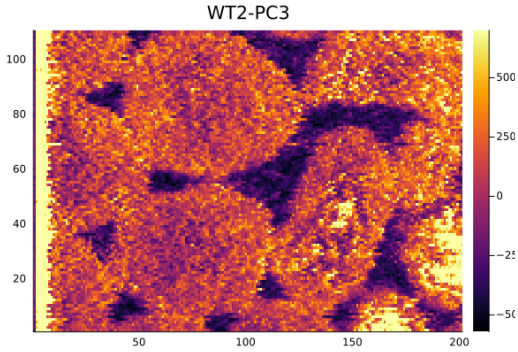
(b) PCA スコアイメージング-KO2-PC2



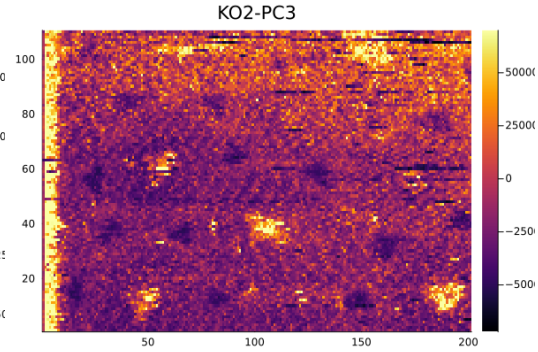
(c) PCA スコアイメージング-WT2-PC2



(d) PCA スコアイメージング-KO2-PC2

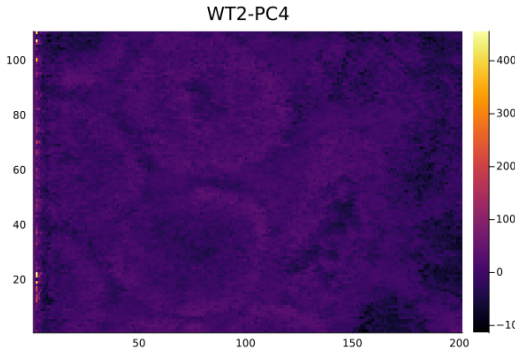


(e) PCA スコアイメージング-WT2-PC3

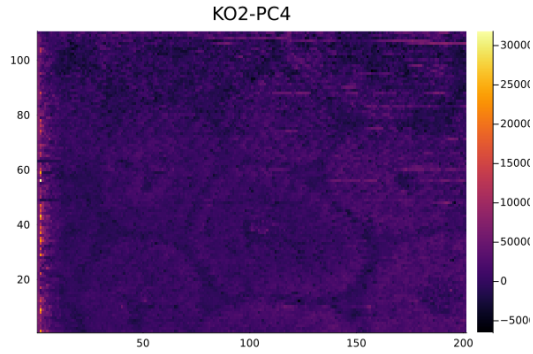


(f) PCA スコアイメージング-KO2-PC3

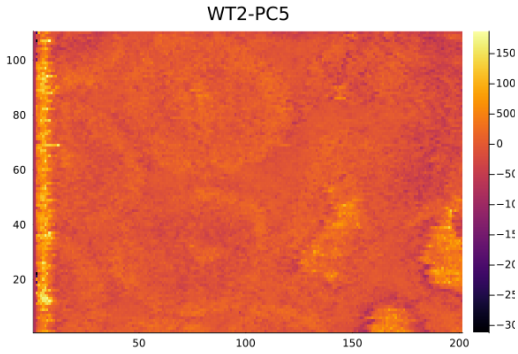
図 48: PCA スコアイメージの比較 (WT2, PC1-PC3)



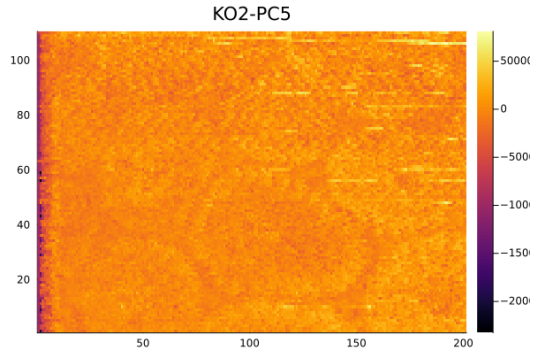
(a) PCA スコアイメージング-WT2-PC4



(b) PCA スコアイメージング-KO2-PC4



(c) PCA スコアイメージング-WT2-PC5



(d) PCA スコアイメージング-KO2-PC5

図 49: PCA スコアイメージの比較 (WT2, PC4-PC5)

7.2 線形判別分析の結果

続いて, 教師あり学習である線形判別分析 (LDA) を行った. 各データ点（全データ 176679, WT: 110349, KO: 66330）に, WT(野生型), KO(ノックアウト型) のラベルを与え, 50 %のデータを学習用に, 残り 50 %のデータを検証用とした. 学習用データを用いて判別関数を構築した際の, 判別関数の重み (MATLAB では DeltaPredictor と表示される) と, 各 m/z の関係をプロットした図を図 50 に示す. ここで, DeltaPredictor は, 各 m/z が判別にどの程度影響しているかを示す指標となっている. この値が大きいとき, 該当する m/z は, 判別に大きく影響していると言える. また, テスト用データを用いた検証結果を混同行列として, 図 51 に示す. テスト

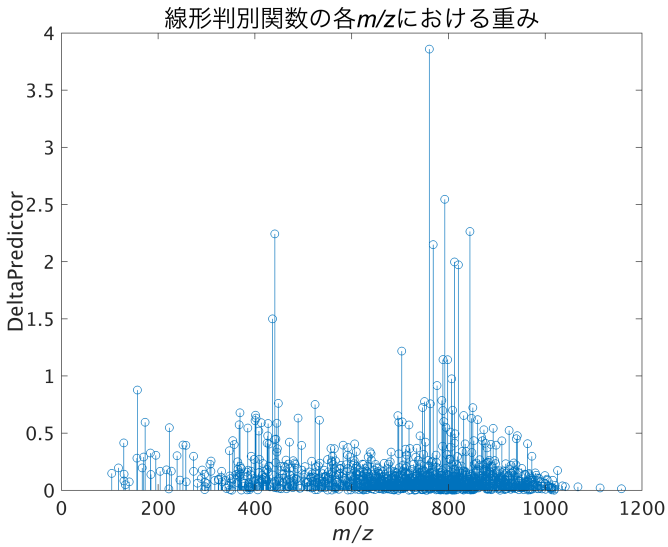


図 50: 線形判別による分析結果. 横軸は m/z, 縦軸は判別関数の重み (DeltaPredictor) を示す.

用データ点 88339 個に対し, 正解数 88303 個 (正解率:99.96 %) という結果が得られた. この結果から, このモデルは検証用データに対して, 高精度で分類できたことを示す.

次に, 線形判別分析によって得られた, 重みの大きい m/z に対して, LIPIDMAPS で帰属された脂質を表 3 に示す. この表において, Name の欄に記載があるものは, その質量に帰属される脂質があったことを表す. Ion-1 の欄は, 脂質が帰属された場合, プロトン付加, ナトリウム付加, カリウム付加のいずれであるかを示す. 一方, Name および Ion の欄に記載がないものについては, その質量に帰属される脂質がなかったことを表している.

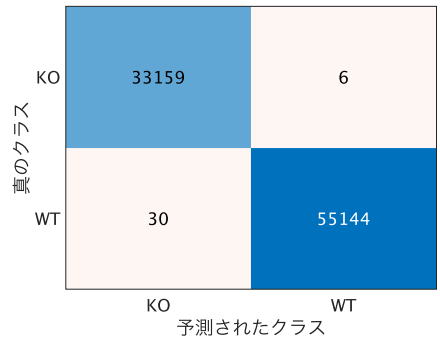


図 51: テスト用データを用いた線形判別モデルの検証結果.

表 3: 線形判別関数の重みの大きい m/z に帰属される脂質の一覧.

m/z	DeltaPredictor	Name-1	Ion-1
760.586	3.85864744	PC 34:1	[M+H]+
792.590	2.54488639	PC O-38:6	[M+H]+
844.526	2.26376076	PC 38:6	[M+K]+
441.240	2.24209274	MG 22:6	[M+K]+
768.590	2.14847663	PC O-36:4	[M+H]+
812.611	1.99755328	PC 36:0	[M+Na]+
820.526	1.97189516	PC 36:4	[M+K]+
436.342	1.49979357		
703.575	1.21754001	SM 34:1;O2	[M+H]+
788.616	1.14267216	PC 36:1	[M+H]+
798.542	1.14221292	PC 34:1	[M+K]+
806.569	0.97517219	PC 38:6	[M+H]+
776.580	0.91581834	PS O-36:1	[M+H]+
157.083	0.87638442	FA 8:2;O	[M+H]+
786.598	0.78581905	PC 36:2	[M+H]+
750.543	0.77536692	PE O-38:6	[M+H]+
448.291	0.76117697		
762.644	0.75660119		
524.371	0.75051562	LPC 18:0	[M+H]+
746.606	0.72398887	PC O-34:1	[M+H]+

m/z	DeltaPredictor	Name-1	Ion-1
850.565	0.72263653		
808.585	0.70065707	PC 38:5	[M+H]+
788.502	0.69797862	PE O-38:6	[M+K]+
369.225	0.67852199	FA 18:1;O4	[M+Na]+
401.266	0.65654697	MG 20:4	[M+Na]+
831.623	0.65432709		
695.539	0.65312179		
402.371	0.63126474		
489.319	0.63092636		
847.615	0.62970323		
860.522	0.61726357	PS O-40:6	[M+K]+
533.347	0.61362582		
400.342	0.60964958	CAR 16:0	[M+H]+
790.571	0.60086306	PC O-36:4	[M+Na]+
704.391	0.59793392		
697.414	0.59500703		
173.078	0.59478723	FA 8:2;O2	[M+H]+
413.266	0.59134472	ST 24:2;O4	[M+Na]+
445.292	0.5868263		
427.282	0.58411156		

m/z	DeltaPredictor	Name-1	Ion-1
367.336	0.57310568		
718.575	0.57149413	PC O-32:1	[M+H]+
794.604	0.55009688	PC O-36:2	[M+Na]+
223.107	0.54728465		
385.272	0.54481119	MG O-20:5	[M+Na]+
892.514	0.54207767		
873.458	0.52962443		
925.520	0.52340636	CL 36:4	[M+H]+
408.368	0.5182595		
804.495	0.50856165	PE 38:5	[M+K]+
813.685	0.4935472	SM 42:2;O2	[M+H]+
942.512	0.47716624		
741.434	0.4761488		
426.358	0.47560231	CAR 18:1	[M+H]+
868.517	0.46425951		
940.497	0.4505366		
443.256	0.44788154		
441.188	0.4437458	FA 20:3;O6	[M+K]+
870.539	0.43459608	PC 40:7	[M+K]+
852.584	0.43407759	SHexCer 38:1;O3	[M+H]+

m/z	DeltaPredictor	Name-1	Ion-1
354.285	0.43354956		
910.523	0.43233172		
752.557	0.42115353	PE O-36:2	[M+Na]+
471.287	0.42072763	ST 28:1;O4	[M+K]+
128.953	0.41449812		
812.529	0.41269715		
425.266	0.41224974	MG 22:6	[M+Na]+
876.518	0.41197383	PS 40:5	[M+K]+
963.477	0.40728138	CL 36:4	[M+K]+
834.598	0.40563839	PC 38:3	[M+Na]+
606.296	0.40536443	LPC 22:6	[M+K]+
429.298	0.40462329	ST 28:3;O4	[M+H]+
887.656	0.40196617	TG 52:7	[M+K]+
357.240	0.39985052		
251.185	0.39626098		
582.296	0.39464658	LPC 20:4	[M+K]+
898.554	0.39412615		
496.340	0.39281365	LPC 16:0	[M+H]+
257.147	0.39258236		
787.669	0.38476401	SM 40:1;O2	[M+H]+

m/z	DeltaPredictor	Name-1	Ion-1
804.609	0.37765123	PS O-38:1	[M+H]+
592.244	0.37418849		
879.620	0.37138721		
734.570	0.36774211	PC 32:0	[M+H]+
720.589	0.36520626	PC O-32:0	[M+H]+
557.095	0.3650406		
564.249	0.36412143	LPE 22:6	[M+K]+
446.244	0.36362303		
711.530	0.36018204	PA O-36:1	[M+Na]+
347.183	0.34618058	FA 18:4;O3	[M+Na]+
836.615	0.34040217	PC 40:5	[M+H]+
828.551	0.34009648	PC 38:6	[M+Na]+
609.524	0.33697158		
638.216	0.33540312		
445.266	0.3350409	LPA O-18:1	[M+Na]+
681.485	0.33399437	PA O-36:5	[M+H]+
848.616	0.33052872	PE 44:6	[M+H]+
184.073	0.3252897		
640.215	0.31930622		
697.478	0.31904703	PA 34:1	[M+Na]+

表 3 に示された m/z の, イオン像を WT, KO 同士で比較したものを図 52 に示す. (a) では, WT で CST 外部のみが強い強度を示し, (c) では CST 内部のみが強い強度を示しているのに対し, (b) では, CST 全体で強い強度を示している. (d) や (f) では, CST ではなく間質とよばれる部分のみが強い強度を示している. (g) や (h) では, WT と KO で明確な違いは確認できず, (i) では CST 全体において, KO より WT が高い強度を示している. 特に注目すべきは, 二重結合を 6 つ含む DHA 含有リン脂質と推定されるもの ((c) m/z 844.526) である. 本計測で用いたマウスは, LPAAT3 という, 生体膜リン脂質生合成酵素の一つである LPAAT3 をノックアウトしたマウスであり, この酵素は特に DHA 含有リン脂質の生合成に関与している. 線形判別関数の重みとして, これらの脂質が大きい値を示していることは, LPAAT3 をノックアウトしたことにより, DHA 含有リン脂質が減少したと読み取れる. また, WT について, 各領域で特徴的な脂質分布を示す m/z 768.589 について, 測定領域ごとの画像を比較した. そのイオン像を図 53 に示す. ここで注目すべきは, LDA で重みが大きいとされた他のイオン像は, WT でイオン強度が強く, KO でイオン強度が低いものであったのに対し, m/z 768.589 については, KO でより強い強度を示している点である.

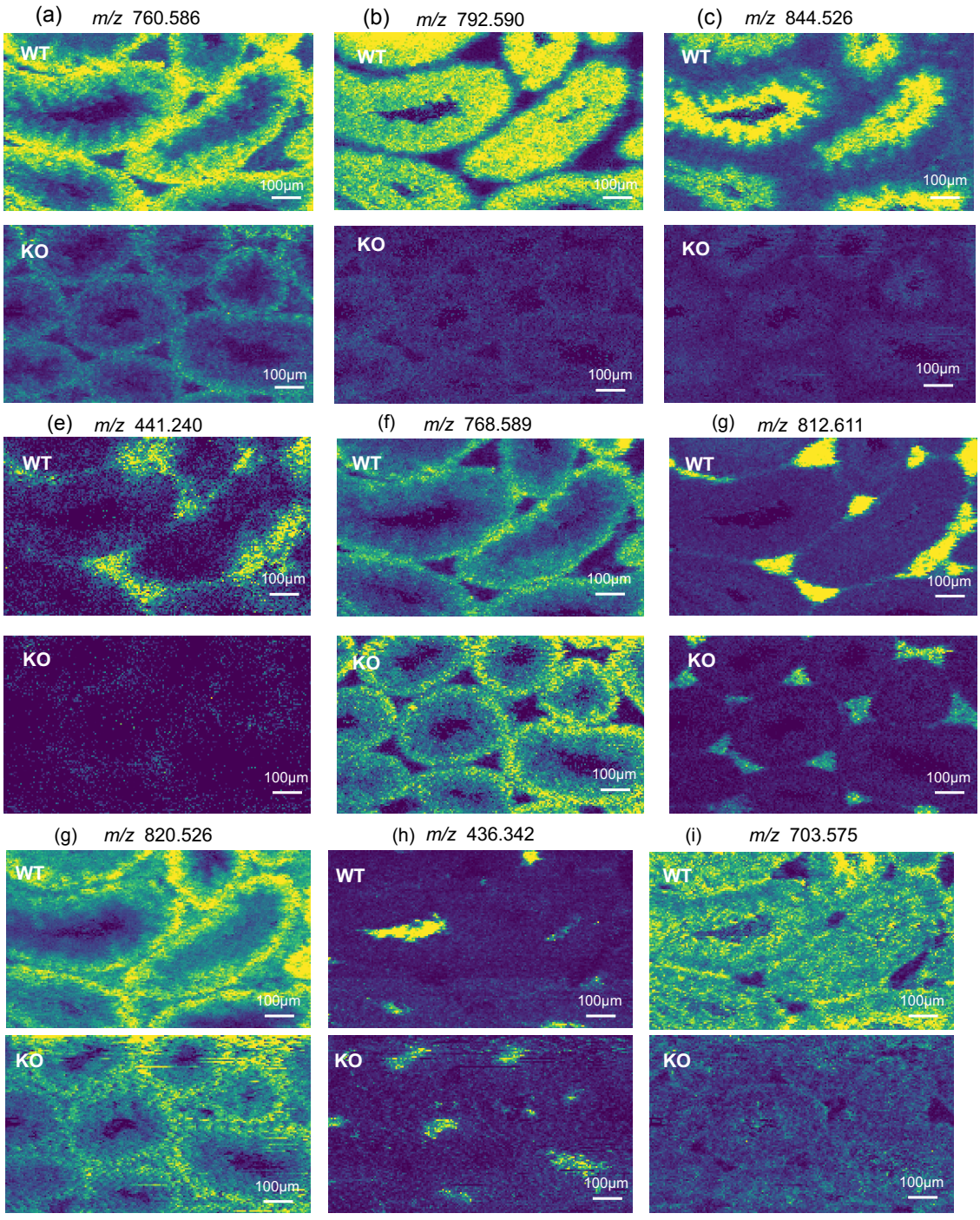


図 52: LDA の判別関数の重みが大きい m/z のイオン像.

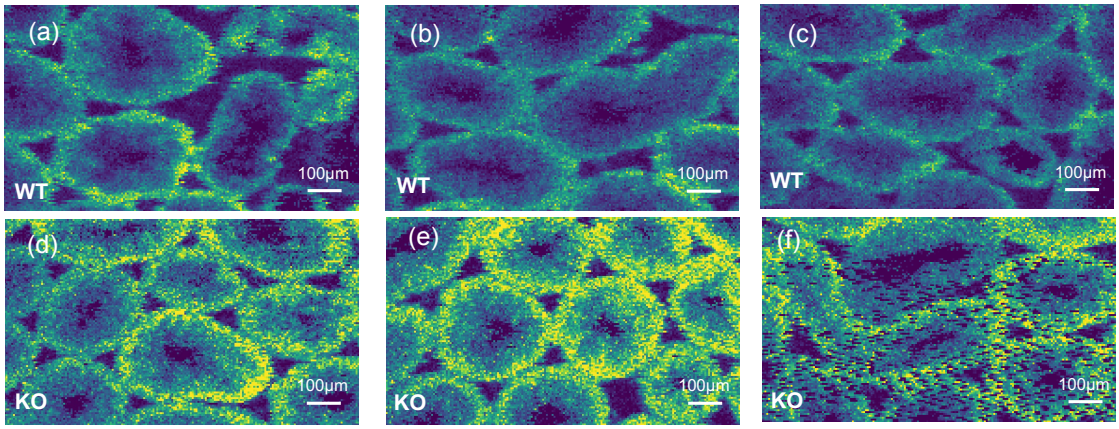


図 53: m/z: 768.589396 のイオン像.

7.3　線形 SVM による分析

次に, 線形サポートベクターマシーン (SVM) を用いて分析を行った. LDA と同様, 各データ点（ピクセル）に, WT(野生型), KO(ノックアウト型) のラベルを与え, 50 %のデータを学習用に, 残り 50 %のデータを検証用とした. 学習用データを用いて判別関数を作成した際の, 判別関数の重み (MATLAB では Beta と表示される) と, 各 m/z の関係をプロットした図を図 54 に示す. ここで, Beta は, 各 m/z が判別にどの程度影響しているかを示す指標となっている. この値が大きいとき, 該当する m/z は, 判別に大きく影響していると言える. また, テスト用

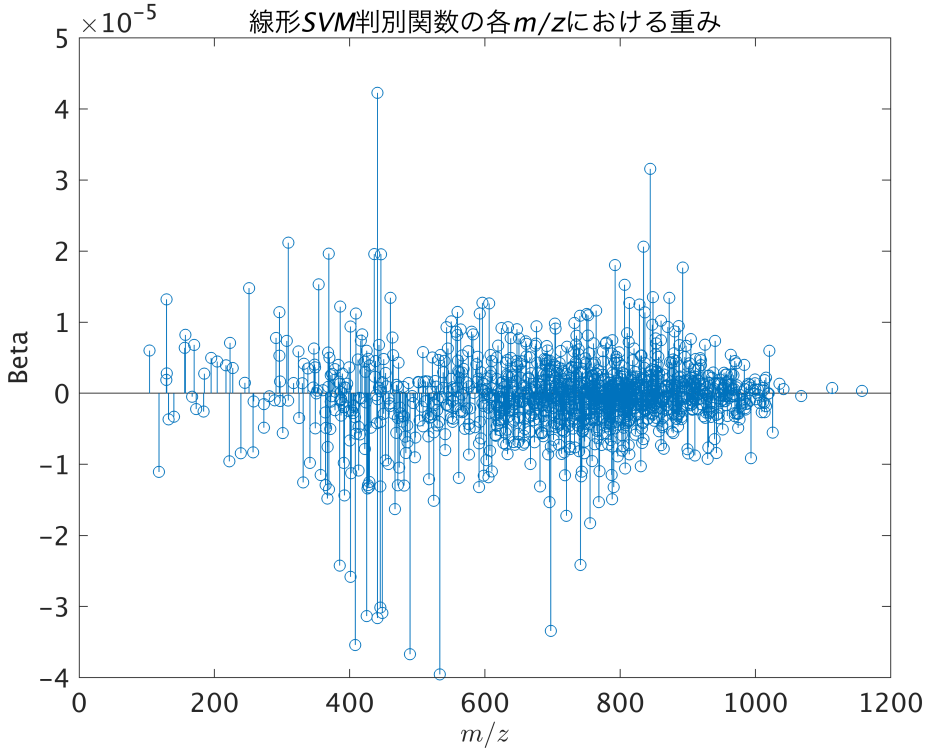


図 54: 線形 SVM による分析結果. 横軸は m/z, 縦軸は判別関数の重み (Beta) を示す.

データを用いた検証結果の混同行列を図 55 に示す. テスト用データ点 88339 個に対し, 正解数 88330 個 (正解率:99.99 %) という結果が得られた. この結果から, このモデルは検証用データに対して, 高精度で分類できたことを示す.

線形 SVM によって得られた, 重みの大きい m/z に対して, LIPIDMAPS で帰属された脂質を表 4 に示す.

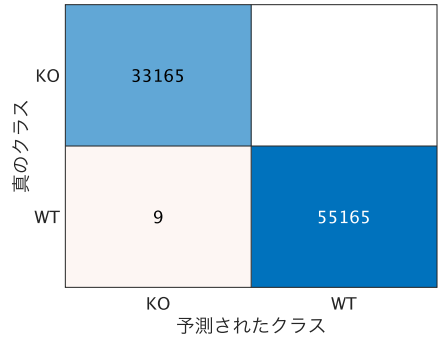


図 55: テスト用データを用いた線形 SVM モデルの検証結果.

表 4: 線形 SVM 判別関数の重みが正に大きい m/z に帰属される脂質の一覧.

m/z	Beta	Name-1	Ion-1
441.240	4.23E-05	MG 22:6	[M+K]+
844.526	3.16E-05	PC 38:6	[M+K]+
309.227	2.12E-05		
834.598	2.06E-05	PC 38:3	[M+Na]+
446.244	1.96E-05		
436.342	1.96E-05		
369.225	1.96E-05	FA 18:1;O4	[M+Na]+
792.590	1.80E-05	PC O-38:6	[M+H]+
892.514	1.77E-05		
806.569	1.53E-05	PC 38:6	[M+H]+
354.285	1.53E-05		
251.185	1.48E-05		
848.616	1.35E-05	PE 44:6	[M+H]+
872.558	1.34E-05	PC 40:6	[M+K]+
460.270	1.34E-05		
129.102	1.32E-05		
813.685	1.27E-05	SM 42:2;O2	[M+H]+
596.355	1.27E-05	PC 20:1;O2	[M+H]+
606.296	1.26E-05	LPC 22:6	[M+K]+
828.551	1.25E-05	PC 38:6	[M+Na]+
386.026	1.22E-05		

m/z	DeltaPredictor	Name-1	Ion-1
764.522	1.16E-05	PE 38:6	[M+H]+
836.615	1.14E-05	PC 40:5	[M+H]+
296.243	1.14E-05		
559.398	1.14E-05		
409.271	1.12E-05	FA 22:1;O2	[M+K]+
592.244	1.12E-05		
750.543	1.11E-05	PE O-38:6	[M+H]+
752.557	1.10E-05	PE O-36:2	[M+Na]+
740.523	1.09E-05	PE 36:4	[M+H]+
860.522	1.02E-05	PS O-40:6	[M+K]+
550.181	1.01E-05		
732.425	9.89E-06		
703.575	9.80E-06	SM 34:1;O2	[M+H]+
847.615	9.65E-06		
886.571	9.45E-06	PE 44:6	[M+K]+
676.358	9.43E-06		
401.185	9.39E-06		
870.539	9.31E-06	PC 40:7	[M+K]+
634.307	9.28E-06	PC 20:1;O2	[M+K]+
543.424	9.28E-06		
624.387	9.17E-06	PS 24:0	[M+H]+
704.391	9.11E-06		
647.450	9.00E-06		
564.249	8.99E-06	LPE 22:6	[M+K]+
880.554	8.69E-06		
580.361	8.67E-06	LPS 22:1	[M+H]+
557.095	8.66E-06		
810.486	8.45E-06		
418.369	8.28E-06		

m/z	Beta	Name-1	Ion-1
533.347	-3.95E-05		
489.319	-3.67E-05		
408.368	-3.54E-05		
697.414	-3.34E-05		
441.188	-3.16E-05	FA 20:3;O6	[M+K]+
425.214	-3.13E-05	FA 20:3;O6	[M+Na]+
448.291	-3.09E-05		
445.292	-3.01E-05		
401.266	-2.58E-05	MG 20:4	[M+Na]+
385.272	-2.42E-05	MG O-20:5	[M+Na]+
741.434	-2.42E-05		
755.503	-1.83E-05	DG 44:10	[M+K]+
720.589	-1.72E-05	PC O-32:0	[M+H]+
467.102	-1.63E-05		
695.538	-1.53E-05		
768.589	-1.53E-05	PC O-36:4	[M+H]+
524.371	-1.51E-05	LPC 18:0	[M+H]+
788.252	-1.49E-05		
367.139	-1.48E-05		
392.373	-1.44E-05		

m/z	Beta	Name-1	Ion-1
369.297	-1.35E-05	WE 22:2;O2	[M+H]+
426.358	-1.34E-05	CAR 18:1	[M+H]+
428.373	-1.32E-05	CAR 18:0	[M+H]+
591.389	-1.32E-05		
790.571	-1.32E-05	PC O-36:4	[M+Na]+
681.485	-1.31E-05	PA O-36:5	[M+H]+
445.267	-1.31E-05	LPA O-18:1	[M+Na]+
480.424	-1.30E-05		
471.287	-1.30E-05	ST 28:1;O4	[M+K]+
427.282	-1.29E-05		
364.342	-1.29E-05		
331.209	-1.25E-05		
429.298	-1.25E-05	ST 28:3;O4	[M+H]+
517.350	-1.21E-05		
561.377	-1.19E-05		
605.403	-1.18E-05		
742.451	-1.17E-05		
598.296	-1.16E-05	PS 20:1;O2	[M+H]+
718.575	-1.15E-05	PC O-32:1	[M+H]+
787.480	-1.15E-05		
357.240	-1.15E-05		
402.371	-1.13E-05		
118.086	-1.11E-05		
743.423	-1.10E-05		
610.371	-1.10E-05		
769.481	-1.10E-05	PA 42:10	[M+H]+
413.324	-1.09E-05		
807.513	-1.06E-05		
473.324	-1.04E-05		
830.561	-1.03E-05		

表 4 に示された m/z のうち, 上位 10 個について, イオン像を WT, KO 同士で比較したものを図 56 に示す. (a)〜(e) は重みが正であったもの（野生型:WT を判別する上で重要な要素）, (f)〜(j) は重みが負であったもの (ノックアウト型: KO を判別する上で重要な要素) である. イオン像からは WT と KO で明確な違いが見られるものの, 特に重みが負であったものについては, バックグラウンドの影響（溶媒が組織成分と混合したなど）が大きく, 脂質によるピークではないと考えられる. よって, 分析としては WT と KO とを分類することができているが, 生物学的な意味は少ないと考えられる.

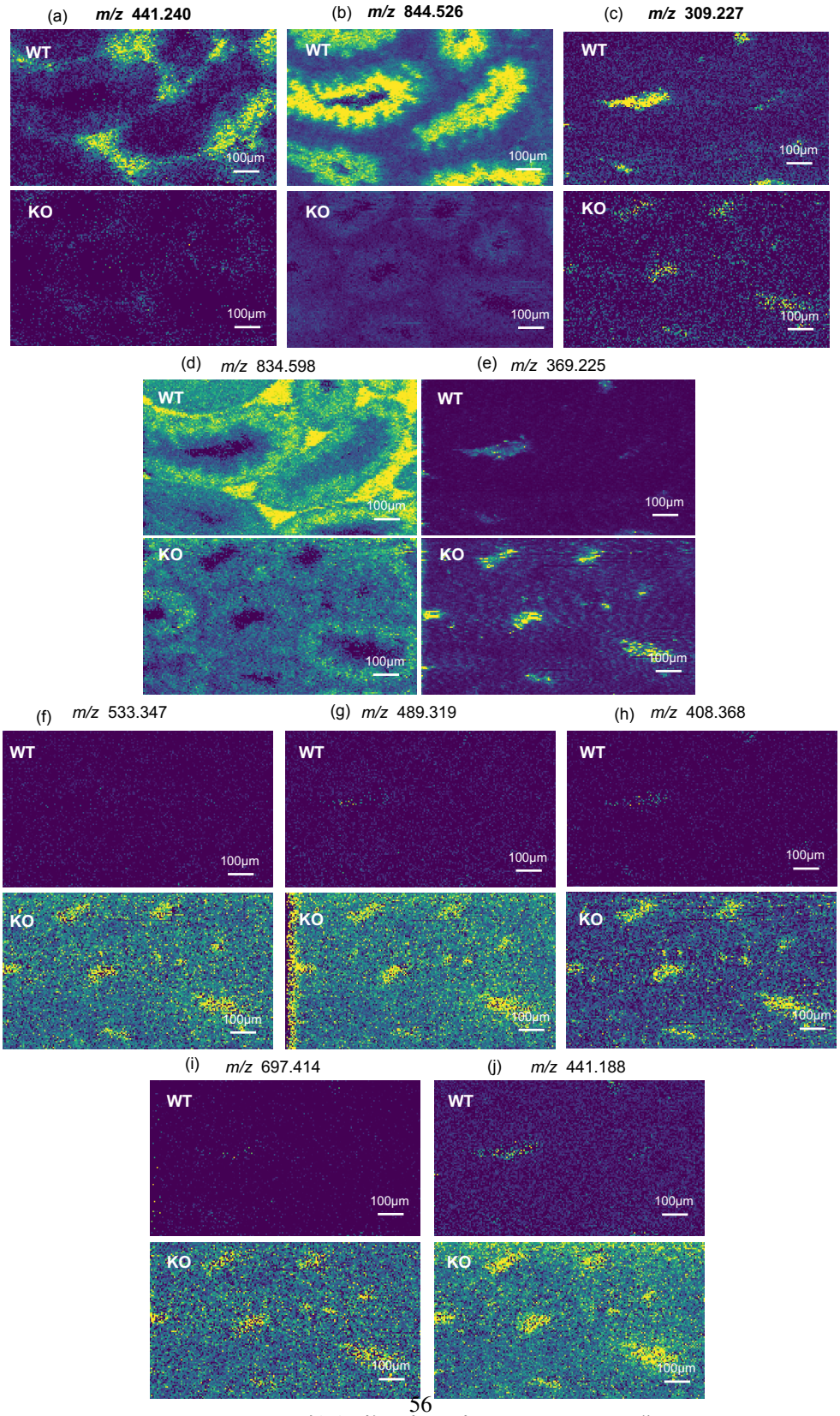


図 56: SVM の判別関数の重みが大きい m/z のイオン像.

8 まとめ

本研究では, 当研究室で取得されたデータ, および自ら取得したデータを用いて機械学習の手法を複数用いて分析を行った. まず, 自ら取得したデータを用いて分析を行なった第 6 章の結果から, 次のようなことがわかった.

1. マウス精巣組織中の CST では, 脂質の分布強度に局在性が見られる.
2. k-平均法といった, 教師なし学習によるクラスタリングによって, CST 内部を 3 種類へと分類することができた. これら 3 種類は, CST 内で階層的に分類されていた.
3. k-平均法によるクラスタリングの分離度を検証するために, t-SNE によるクラスタリングを行なった結果, k-平均法による 3 種類へのクラスタリングは, 各クラスターは十分分離されていた. しかし, 4 種類へのクラスタリングでは, クラスター同士の重なりが確認できたため, 十分に分類されているとは言えない.
4. このようなクラスタリングをもとに, 各データ点（ピクセル）に対しラベル付けを行い, 教師あり学習を行うことで, 局在する脂質を特定することが期待できる.

次に, 当研究室で取得されたデータを用いて分析を行った第 7 章の結果から, 次のようなことがわかった.

1. マウス精巣組織の WT と KO を PCA により比較すると, ローディングベクトルとスコアイメージングから CST 内部の脂質分布に違いがあることはわかったが, どの m/z で有意な差があるかまでは判明しなかった.
2. 教師あり学習である線形判別分析, 線形 SVM の双方で, WT と KO を 100 %に近い形で分類することが可能であった.
3. 線形判別分析, 線形 SVM の双方で, 判別関数の重みから判別に重要な意味を持つ m/z を導くことができた. 特に, 線形判別分析で重み, DeltaPredictor の数値が大きいものは, イオン像で確認すると CST 内の脂質分布に明確な違いがあり, 生物学的に重要な脂質が多く含まれていた.

従来は, 生体試料を提供していただいた方からの事前情報や, 各 m/z の画像を目視で野生型 (WT) とノックアウト型 (KO) で脂質分布に差があるかどうかを確認していたため, 分析に非常に多くの時間を必要としていた. また, 差の有無を客観的に評価する方法が乏しい状況であった. 今回用いた方法では, 短時間で WT と KO とで違いがあり, かつ生物学的に意味のある m/z を導き出すことができただけでなく, 多変量解析による分析を行ったことで, WT と KO に違いがあるという結論により客観性を持たせることができた. また, マウス精巣内の脂質の強度分布について, 今後, 各 m/z のイオン像に対して画像分析を行うことで, より詳細に, マウス精巣組織内の脂質分布の違いを導くことが期待できる.

謝辞

豊田岐総先生, 兼松泰男先生には, 貴重なご意見ご指導賜った.
大塚洋一先生には, 論文の添削, 発表スライドの添削, 文章の書き方等貴重なご意見ご指導賜った.
河井洋輔先生には, 貴重なご意見ご指導賜った.
国立国際医療研究センターの進藤　英雄先生には, 本研究で用いたサンプルを提供していただいた. 名古屋市立大学データサイエンス学部教授　小山　聡　先生には, 機械学習に関するご指導およびアドバイスをいただいた. 当研究室 D1 孫さんには, 実験方法について助言をいただいた. M2 岡田さんには, 実験方法の指導, および 7 章で用いたマウス精巣組織データを提供していただいた. その他, 豊田研究室の学生の皆様に多方面でお世話になった. ここに感謝の意を表する.

付録 A　MATLAB の導入

本研究では, 取得したデータを分析する際に, 数値解析ソフトウェアである MATLAB を利用した. MATLAB は他の言語と比較し, 短時間かつ簡単に科学技術計算ができ, その結果の可視化にも長けている. また, 専用の拡張パッケージ (Toolbox) を導入することで, より高度な分析や操作が可能となる.
(MATLAB; MathWorks https://jp.mathworks.com/products/matlab.html)

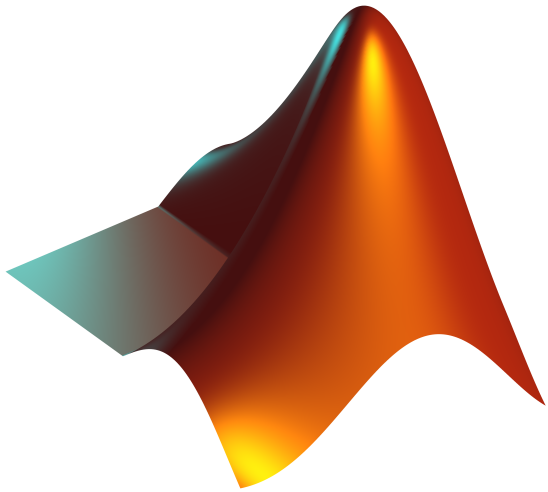


図 57: MATLAB ®

A.1　MATLAB ライセンス

本研究の分析では, 基本的に自らの PC で分析を行った. MATLAB の学生向けライセンス”MATLAB and Simulink Student Suite”を購入し,
(https://jp.mathworks.com/products/matlab/student.html) このライセンスに付属しているパッケージで多くの分析は実行可能であった. 特に, 統計解析, 機械学習による分析を行う際は, ”Statistics and Machine Learning Toolbox”が導入されていれば, 多くの分析方法が 1 つのコードで実行できるため, 便利である.

A.2　MATLAB へのデータダウンロード

続いて, MATLAB へのデータダウンロード方法について記述する. MATLAB へデータをインポートする方法は大きく分けて 2 つあり,

MATLAB
Simulink
Control System Toolbox
Curve Fitting Toolbox
DSP System Toolbox
Image Processing Toolbox
Instrument Control Toolbox
Optimization Toolbox
Parallel Computing Toolbox
Signal Processing Toolbox
Statistics and Machine Learning Toolbox
Symbolic Math Toolbox

表 5: MATLAB and Simulink Student Suite に付属しているパッケージ.

(i) **「データのインポート」ボタンからデータをインポートする.**

(ii) **コード入力によりデータをインポートする.**

(i) の方法では, 現在のディレクトリ以外の場所からでもデータをインポートすることができる. (ii) の方法では, 現在のディレクトリにあるファイルを, 任意の変数名としてインポートすることができる. データをインポートする際には, table 型（表, 変数行がついている）, double 型 (数値行列), string 型（文字配列）など, 元のデータの性質やインポートしたい形を考え, 導入することができる.

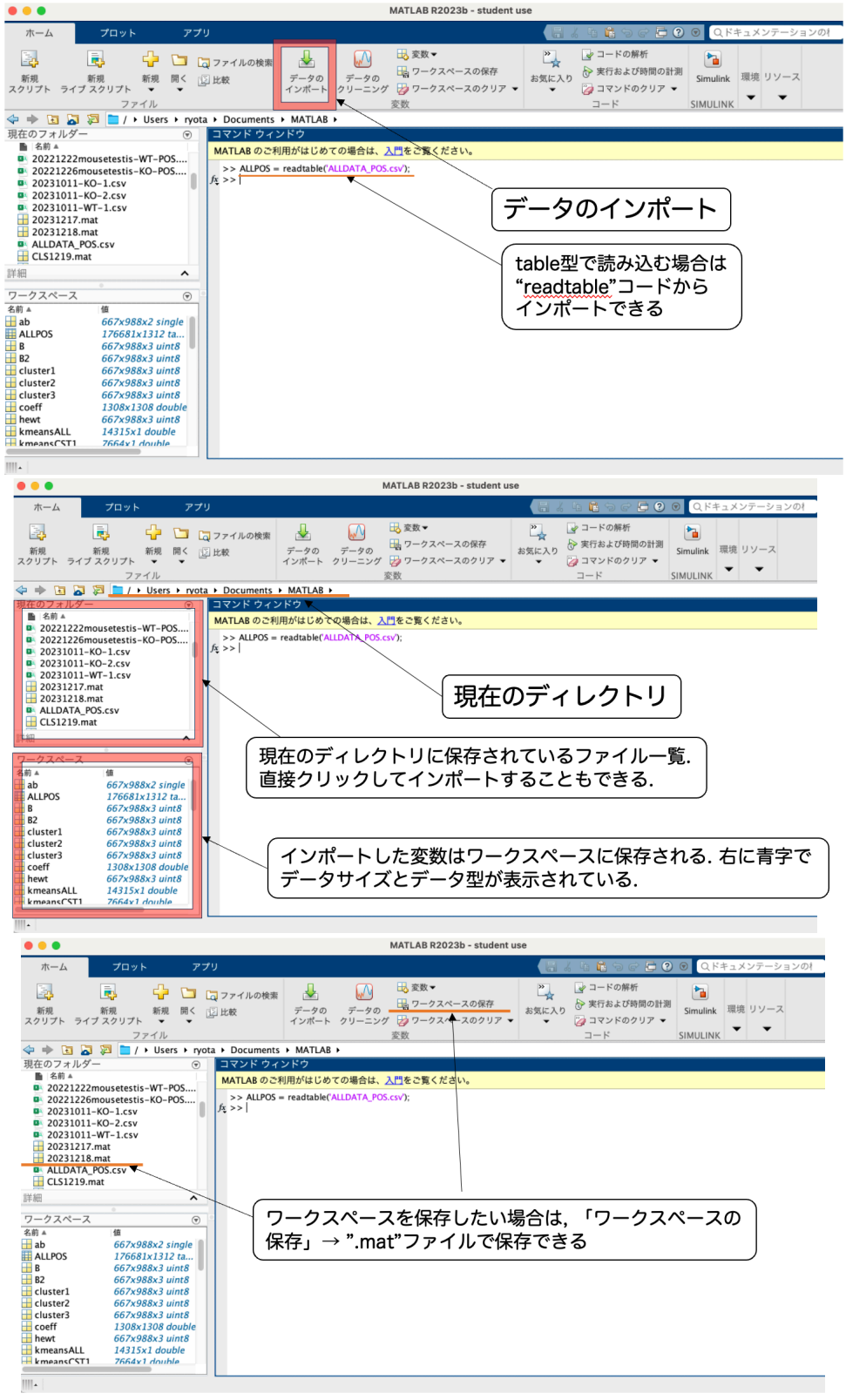


図 58: MATLAB の操作画面

A.3　MATLAB での各種分析の説明

A.3.1　データの出力方法

IMAGEREVEAL にてデータを出力するには, 差異解析を行ったプロジェクトファイルを開いた後, 任意の IMDX ファイル名を右クリックし, 「データ行列を出力」をクリックする. その後, 保存したい場所とファイル名を指定すると, CSV 形式で出力が開始される. 出力されたデータは Microsoft Excel で開くと, 図 59 のような形式となっている. WT と KO の判別を行う場合には, WT と KO のデータを 1 つのファイルに結合しておく.

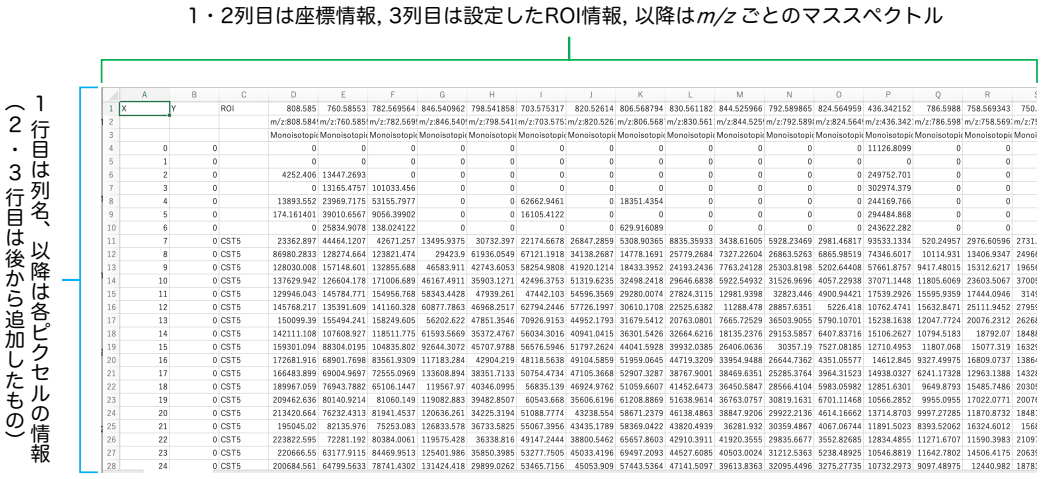


図 59: IMAGEREVEAL から出力したデータの例. 2 行目と 3 行目は後から追加したものであり, 元のデータにはない.

A.3.2　PCA

MATLAB にて PCA を行うには, 次のような手順をとる. まず, インポートしたデータが table 型（列名がついている）であった場合, まず, 数値行列 (double 型) へと変換する必要がある. table 型→ double 型への変換は, コマンドにて

(**変数**) = table2array(**変数**)

と入力する. 元のデータに, X, Y, ROI など解析に必要のない列やタイトル行などが含まれている場合は, あらかじめ削除しておく. 削除した後のデータは, 各 m/z のマススペクトルがピクセル行分並んでいるもの (ピクセル数が m, m/z の数が n とすると, m×n のデータ) になっている. このようにして, データを編集した後, pca() というコマンドを用い, PCA を行う. PCA の結果については, 本研究では主成分係数 (負荷量), スコア, 寄与率を用いたため, コマンドに

[coeff,score,latent] = pca(**変数**)

と入力し, 結果を得た. ここで, coeff=主成分係数, score=主成分スコア, latent=寄与率である. coeff と latent は (m/z の数)×1 の行列, score は (ピクセル数)×(主成分の数 (m/z の数)) の行列として得られる.

分析結果をプロットするには, コマンドにて

figure
plot(m/z の一覧, latent(:,1))　または　stem(m/z の一覧, latent(:,1))

と入力する. 入力内容によって表示されるまでの時間に違いはあるが, 概ね 30 秒以内に表示される. plot() の場合は折れ線グラフ, stem() の場合は茎グラフ (図 60) で表示される. その他にもさまざまなグラフオプションが用意されているので, それを利用したい場合は画面上の”プロット”タブで任意のグラフ形式を選択されたい.

また, データセットの 1 列あるいは 1 行のみを選択する場合, データセットの名前を X とすると,

X(:,1)(**1 列目のみを抜き出す場合**)　X(1, :)(**1 行目のみを抜き出す場合**)

のように入力する.

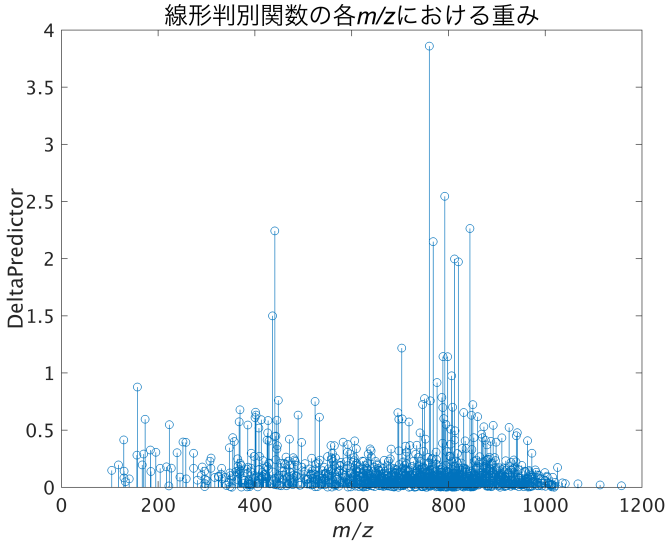


図 60: stem グラフの一例.

A.3.3 k-means clustering

MATLAB で k-means クラスタリングを行う場合は, PCA の際に利用したものと同様のデータセットを用い,

```
kmeans(変数)
```

のように入力する. すると, 結果として行：ピクセル数 × 列：ラベルのデータセットが出力される.

A.3.4 t-SNE

MATLAB で t-SNE クラスタリングを行う場合は, PCA の際に利用したものと同様のデータセットを用い,

```
tsne(変数)
```

のように入力する. すると, 結果として行：ピクセル数 × 列：計算結果のデータセットが出力される. この分析結果を散布図に示したい場合, 座標情報と出力されたデータセットを合わせたものを `X`, 使用するラベルを `Y` とすると

```
figure
gscatter(X, Y)
```

と入力すれば, ラベルごとに色分けされた散布図が出力される (図 61).

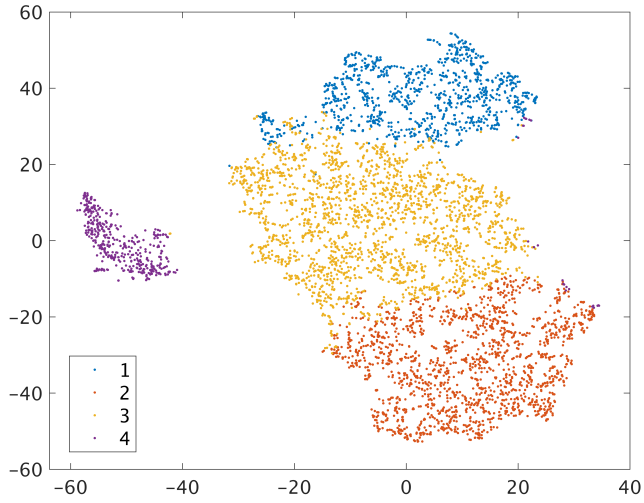


図 61: t-SNE による分析結果を散布図にプロットした例.

A.3.5　判別分析

MATLAB にて判別分析を行う場合は, あらかじめ元のデータから座標情報の列 (X, Y) と ROI 情報の列を削除しておく. 加えて, 新たに WT と KO のラベルを付した行を追加しておく. その後, コマンドにて

`Mdl = fitcdiscr(`**変数，判別に利用するラベルの列名**`)`

と入力する. 分析終了後, "Mdl"変数の中に各種分析結果が表示される. 判別関数の重みを得たい場合は, "`Mdl.DeltaPredictor`"と入力すれば表示することができる.

A.3.6　SVM

線形 SVM での分析を行う場合は, 判別分析を行なったデータと同様のデータセットを用い, コマンドにて

`Mdl = fitcsvm(`**変数，判別に利用するラベルの列名**`)`

と入力する. 分析終了後, "Mdl"変数の中に各種分析結果が表示される. 判別関数の重みを得たい場合は, "`Mdl.Beta`"と入力すれば表示することができる.

A.3.7　その他の分析法を利用したい場合

その他にも MATLAB には, 教師あり学習の手法が多数用意されており, "分類学習器"というアプリケーションを利用することで, 一度に複数の手法による分析が可能となる. 分類学習

器は, MATLAB の”アプリ”一覧から選択する, あるいはコマンドに”classificationLearner”と入力することで起動できる. 分析したいデータセットを選ぶ際に, 検証用にデータを保存しておきたい場合は, テストデータとして任意の割合をランダムで学習とは別に保存しておくこともできる. 分析後, ”結果の解釈”画面で”混同行列”を選択すれば, テストデータの評価を混同行列として示すことができる.

参考文献

[1] 豊田岐総. 質量分析学・基礎編. 日本質量分析学会, 2016.

[2] Yoichi Otsuka, Sayuri Shide, Junpei Naito, Masafumi Kyogaku, Hiroyuki Hashimoto, and Ryuichi Arakawa. Scanning probe electrospray ionization for ambient mass spectrometry. *Rapid Communications in Mass Spectrometry*, Vol. 26, pp. 2725–2732, 12 2012.

[3] Guillaume Robichaud, Kenneth P. Garrard, Jeremy A. Barry, and David C. Muddiman. Msireader: An open-source interface to view and analyze high resolving power ms imaging files on matlab platform. *Journal of the American Society for Mass Spectrometry*, Vol. 24, pp. 718–721, 5 2013.

[4] 日本生化学会. 脂質 II, リン脂質. 東京化学同人, 2022.

[5] A.Kornberg J. Wittenberg. *Journal of Biological Chemistry*, Vol. 202, pp. 431–44, 5 1953.

[6] S.B. Weiss E.P. Kennedy. *Journal of Biological Chemistry*, Vol. 222, pp. 193–214, 9 1956.

[7] Andrew S. Mason, Claire L. Varley, Olivia M. Foody, Xiang Li, Katie Skinner, Dawn Walker, Tony R. Larson, Daisuke Wakamatsu, Simon C. Baker, and Jennifer Southgate. Lpcat4 knockdown alters barrier integrity and cellular bioenergetics in human urothelium. *International Journal of Molecular Sciences*, Vol. 23, , 10 2022.

[8] 毛利秀雄. 精子の生物学. 東京大学出版会, 1991.

[9] Hideo Shindou, Hideto Koso, Junko Sasaki, Hiroki Nakanishi, Hiroshi Sagara, Koh M. Nakagawa, Yoshikazu Takahashi, Daisuke Hishikawa, Yoshiko Iizuka-Hishikawa, Fuyuki Tokumasu, Hiroshi Noguchi, Sumiko Watanabe, Takehiko Sasaki, and Takao Shimizu. Docosahexaenoic acid preserves visual function by maintaining correct disc morphology in retinal photoreceptor cells. *Journal of Biological Chemistry*, Vol. 292, pp. 12054–12064, 7 2017.

[10] Nico Verbeeck, Richard M. Caprioli, and Raf Van de Plas. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews*, Vol. 39, pp. 245–291, 5 2020.

[11] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-sne, 2008.

[12] 藤原幸一. スモールデータ解析と機械学習. オーム社, 2022.

[13] 小西貞則. 多変量解析入門. 岩波書店, 2010.