全ゲノム配列情報と画像クラスタリング技術の融合による希少難治性疾患の層別・分類手法の開発

川 口 修 治

京都大学大学院医学研究科 京都大学・マギル大学ゲノム医学国際連携専攻

１．はじめに

希少難治性疾患（難病）は各疾患の患者数は少ないが、疾患の種類が非常に多いため、全ての患者を合わせると日本における有病率は 8%と言われている。難病は症例数が少ないために、疾患の十分な研究が進まず、明確な診断基準の確立や疾患の分類や定義が不明瞭なままの疾患も多い。一方、多くの難病において、遺伝的要因が発症と関連していることが知られており、発症の要因となる遺伝子（原因遺伝子）の同定が発症機序の解明に重要となっている。ゲノム解読装置の高性能化により、全ゲノム配列を高精度に解読できるようになり、多くの疾患で原因遺伝子が同定されてきた。しかしながら、ゲノムは個人毎に至る所で変異しており、一般的にゲノム解析には十分な数の症例を必要とする。そのため、症例数の少ない難病データに common disease 等で培われてきたゲノム解析手法を適用することが困難である。

このような問題に対し、ゲノム情報に加えて様々なデータを統合的に解析することで原因遺伝子と疾患との関連をより明確に捉えることが期待できる。中でも診断画像は、患者の症状を直接的に表しているため、大きな情報量を持つ。そこで本研究では、診断画像から特徴を抽出し、これをゲノム情報と関連づけることで、原因遺伝子と疾患との関連を明らかにすることを目的とする。特に、本研究では、深層学習を用いて原因遺伝子による診断画像の特徴を分類する手法の開発を目指した。開発は、指定難病の一つである網膜色素変性(指定難病 90)の超広角蛍光眼底写真（広角眼底写真）を用いて進めた。網膜色素変性は網膜に色素変性が起こることで、夜盲や視野狭窄、視力低下が見られる進行性の疾患である。メンデル遺伝病の一つであり、100 近くの原因遺伝子が知られているが[1]、これら既知の原因遺伝子情報を基に、患者のゲノム診断を行っても原因遺伝子が同定できる割合は 3 割を超える程度である[2]。これには、未知の原因遺伝子の存在が原因の一つと考えられるが、他にもゲノムは個人毎に様々な箇所で特異的に変異するため、ゲノム情報のみではどれが真に網膜色素変性に関わる変異であるかを特定することが困難であることも挙げられる。

一方、原因遺伝子の種類が多いことからも、同一の疾患であっても網膜色素変性の病態は、遺伝子や変異によって異なる特徴を持つ可能性が高い。とりわけ広角眼底写真は、網膜色素変性の病態を直接的に表すことから、その特徴の違いがより明確であると考えられる。そこで、網膜色素変性の患者の広角眼底写真を深層学習により学習し、全ゲノム情報から得られた原因遺伝子情報を基に、遺伝子間の違いを分類するクラスタリング手法の開発を行う。本研究で開発するクラスタリングを通じて、難病の詳細な層別・分類を可能とすることで、遺伝子診断率の向上や疾患発症機序の解明を図る。

２．方法

本研究では、256 名の網膜色素変性患者の縦 2,728 ピクセル、横 3,082 ピクセルの広角眼底写真計 468 枚に加えて、健常者の広角眼底写真 104 枚を解析に用いた。256 名の患者の内、168 名については全ゲノム解析によって原因遺伝子が特定されている。はじめに、網膜色素変性画像から 256×256 ピクセルの領域を無作為に 1,400 枚切り出し、これを専門医によって色素変性領域と正常領域に分類した。これを、教師画像として深層学習により学習を行い、色素変性領域と正常領域を分類する識別器を作成した（図１）。次に、網膜色素変性の全体画像に対して、256×256 ピクセルの領域をスライドさせながら切り出して、作成した識別器を適用し、色素変性領域を取り出す。この際、画像の一部が撮影範囲外（黒領域）を含むものは除外する。また、健常者の画像から同様に 256×256 ピクセルの領域をスライドさせながら切り出し、これを正常領域とする。以上で得られた、色素変性領域と正常網膜領域の画像を教師画像として深層学習を再度行い、識別器を更新する(図 1)。このプロセスを繰り返すことで、色素変性領域抽出識別器の精度向上を行う。深層学習には、TensorFlow(version 1.4.0)[3]と Keras(version 2.1.5)[4]を使用し、75%を訓練データとし、残りの 25%を検証データとした。

作成した識別器を用いて、広角眼底写真 314 枚から色素変性領域を抽出する。色素変性領域を 256×256 より細かい解像度で抽出するため、識別器を 64×64 画素ごとにスライドして適用し、周辺領域の色素変性領域の判定値[0-1]（高いほどより色素変性である可能性が高い）を用いて、128×128 の分画で色素変性領域の可能性が高い箇所を抽出し、抽出画像から 128×128 サイズの識別器を作成する（図 2)。次に、得られた 128×128 ピクセルサイズの画像を用いてクラスタリングを行う。クラスタリングには、Metric Learning による深層学習を用いた[5]。抽出した画像には重複部分も多く、学習に偏りが出ることを避けるため、重複部においては識別時に得られる判定値[0-1]が最も高い画像を選別した。はじめに、各画像に対して全ゲノム解析によって得られた、原因遺伝子をクラス ID として付与する。クラス ID を基に Metric Learning により画像の特徴分類を行い、網膜色素変性の特徴が遺伝子間で異なるか調査した。ここで本研究では症例数の多い３遺伝子に限定し、遺伝子のペア毎にクラスタリングを実施し、結果の比較を行なった。

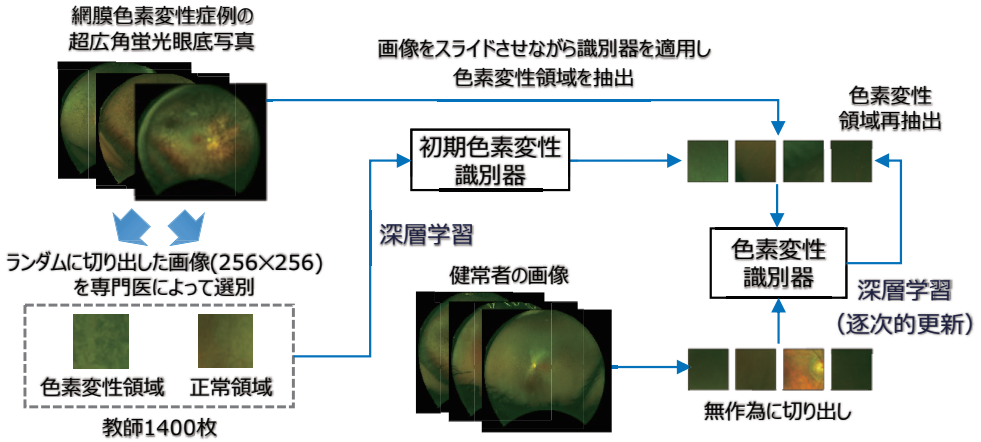


図１： 深層学習による色素変性領域識別器の作成。

専門医によって選別した、教師データを基に作成した色素変性領域識別器を用いて、超広角蛍光眼底写真から色素変性領域を抽出する。抽出した領域の画像と健常者の正常網膜画像から色素変性識別器を再作成する半教師付き学習を行う。

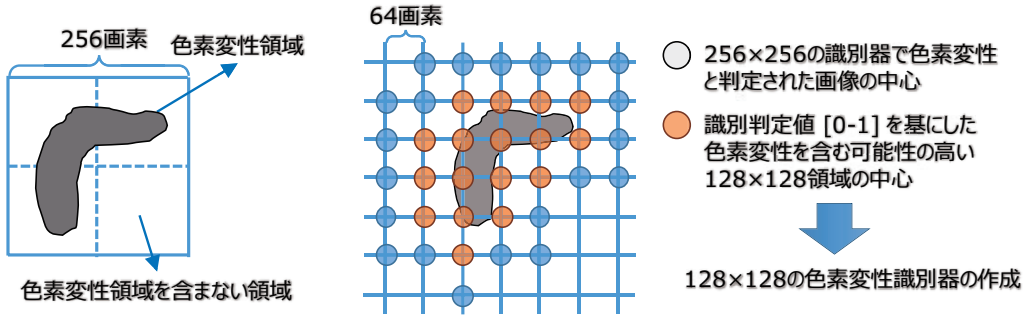


図 2: 色素変性識別器の高解像度化

256×256 の画像は、人手で判定する際には適切なサイズであるが、色素変性パターンの詳細な解析を行うにはサイズが大きい。そこで、256×256 画像の識別器を 64×64 の幅でスライドさせながら適用し、その結果の空間的な識別判定値から、色素変性領域を含んでいる可能性の高い 128×128 画像を選出し、識別器を再作成する。

3．結果

図 3 に、本研究で構築したシステムの詳細を示す。本システムは、仮想コンテナ型システム上で動作するよう設計されており、上位アプリケーションは今後の拡張・統合を見据えて汎用的なプログラムで開発されている。初めに、網膜色素変性の超広角蛍光眼底写真から専門医によって人工的に教師画像を 1,400 枚選別し、初期色素変性識別器を深層学習により作成した。得られた識別器を用いて、網膜色素変性の広角眼底写真 468 枚から 256×256 サイズの色素変性領域を抽出し、健常者の広角眼底写真から切り出した正常領域の画像と併せて約 100 万枚の画像を用意し、これを用いて識別器を逐次的に更新した。ここで、健常者の画像 104 枚のうち 26 枚はテスト用として分離し、識別器の学習には用いなかった。最終的に得られた識別器を改めて広角眼底写真に適用し、64×64 で位置をスライドさせて色素変性の判定を行なったところ、256×256 サイズの色素変性画像が 2,403,625 枚得られた。次に、領域の解像度を上げるため、抽出位置における空間的な判定結果を基に、色素変性箇所が含まれている可能性の高い 128×128 サイズの色素変性領域を 4,071,511 枚抽出し、これらの画像から 128×128 領域用識別器を作成した。識別器の性能をテストするため、26 枚の健常者のテスト画像から 128×128 サイズの画像を 522,138 枚抽出し、色素変性領域の判定を行なった結果、偽陽性率は 0.49%(2,548/522,138)であった。色素変性領域の識別器の学習は Reedbush-H システムの GPGPU2 枚を用いて行い、実行時間は 60 時間を要した。

得られた 128×128 サイズの色素変性画像のうち、頻度の高い３遺伝子で診断された網膜色素変性の広角眼底写真に限定し、Metric Learning によるクラスタリングを行なった。重複部の画像を識別時に得られる判定値を用いて選別し、47,845 枚の画像をクラスタリングに用いた。ここで、病態の進行による影響を除くため、広角眼底写真内で選別された色素変性画像の枚数が 100 枚に満たない症例は除いた。また、画像のうち 50%を教師データ、25%を検証データ、残りの 25%をテストデータとした。図 4 は、ある二つの遺伝子のクラスタリングで得られた、色素変性パターンの特徴空間に対して、テストデータを写像した結果を示している。クラスタリングにより、色素変性画像は Gene A と Gene B の二つの大きなクラスタを形成する結果となった。各クラスタの重心に近い画像を調査したところ、Gene A 側では色素変性が線形的に広がっているのに対して、Gene B 側ではより円状に広がっているパターンが多い結果となった。更に、クラスタの重心

との cosine 類似度を用いて、6634 枚のテスト画像がどちらのクラスタに所属するか分類したところ、Gene A の画像の 95.2%が Gene A のクラスタに、Gene B の画像の 81.7%が Gene B のクラスタ付近に集積していた。Gene B については、画像の 13.9%が Gene A のクラスタ側に所属しており、一部の症例では Gene A に近い色素変性パターンを持つ可能性があった。

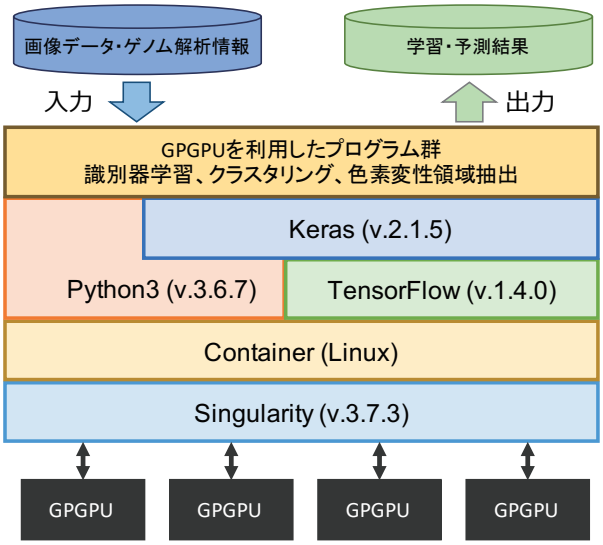
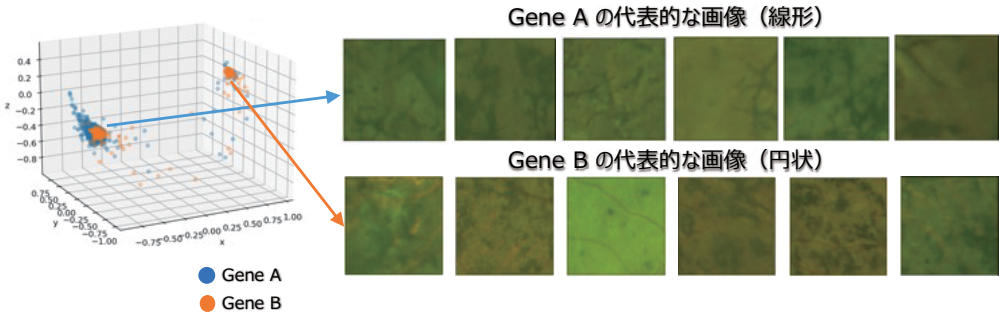


図 3:構築したシステムの構成

Reedbush-H 上で実行した本システムの構成図を示す。深層学習に必要なアプリケーションは仮想コンテナシステムに搭載され、Singularity 上で動作する。



cosine類似度によるテストデータの分類(縦＝予測、横＝実測)

	Gene A	Gene B	外れ値
Gene A	5036	92	164
Gene B	194	1096	52

図 4:二つの原因遺伝子間におけるクラスタリング結果

クラスタリングで分類された二つの原因遺伝子の色素変性領域のテストデータ画像(左上図)。左側に Gene A、右側に Gene B がそれぞれで大きなクラスタを形成している。各クラスタの重心に近い画像を調査した結果、Gene A のクラスタには色素変性が線形的な画像が多く含まれ、Gene B のクラスタには色素変性が円状となる画像が多い結果となった。(右上図)。クラスタ重心との cosine 類似度を用いてテストデータを分類した結果、多くの画像は、同一の原因遺伝子クラスタに含まれたが、Gene B の一部の画像は Gene A のクラスタに含まれた（下表)。

4. 考察

深層学習技術の発展により、筋病理のような専門医でも見分けが困難な難病の診断画像に対してでも、疾患の分類を高精度に行えるようになった[6]。そこで、開発されてきた技術を同じく難病の一つである網膜色素変性の超広角蛍光眼底写真に対しても応用した結果、十分に適用可能であると考えられる。網膜色素変性の原因遺伝子は非常に多く存在するため、その種類によって病態や進行が多岐に渡り、そのために適切な治療法が異なる可能性もある。そのため、原因遺伝子を基準とした、より詳細な疾患の分類が求められる。筋病理の先行研究のように、予め分類すべき疾患が定められているケースでは、各疾患にクラスを割り当てることで分類器を深層学習により作成することが可能であるが、網膜色素変性では、どの遺伝子群が近しいクラスとして属しているのかは不明なため、クラス分類のような深層学習を直接適用できない。本研究では、この問題に対して Metric Learning による深層学習を通じて原因遺伝子間の色素変性パターンの違いを導き出すクラスタリング手法の開発を行なった。初めに、広角眼底写真から色素変性領域を抽出する分類器を作成したが、この時に色素変性領域の抽出と再学習を半教師的に繰り返すことで、専門医による教師データ作成労力を削減した。また色素変性領域抽出の高解像度化により、詳細に色素変性パターンを捉えることを可能とした。

開発したクラスタリング手法を広角眼底写真に適用した結果、二つの原因遺伝子間でパターンの異なる色素変性を検出することができた。今後、これらの遺伝子間の機能的な特徴の違いを調査することで、疾患再分類における有益な知見となることが期待できる。今回、クラスタリング手法の検証を目的とするため、症例数が多い原因遺伝子に限定した。しかしながら原因遺伝子の多くは、一つのクラスにおける教師データとして用いるには困難なほど症例数が少なく、加えて原因遺伝子が未だ不明な症例も多く存在する。本手法は、テストデータによる検証で用いたように、分類器を新規データに適用することで、各画像を特徴空間上に写像することができる。そこで、初期のクラスタリングでは症例数が多い原因遺伝子に限定し、そこで得られた分類器を基に、逐次的に原因遺伝子を追加するなどの工夫を用いて、クラスタの拡張を行う手法の開発を行なっている。この際には、２クラス分類器をより多クラスの分類器へ拡張することも同時に進めている。これにより、網膜色素変性の様々な原因遺伝子群を、類似する色素変性パターン毎に分類していくことを目指す。

謝　辞

本研究は、2020 年度東京大学情報基盤センター若手・女性利用者推薦制度の支援により実施いたしました。本研究の遂行に当たりましては、京都大学医学研究科附属ゲノム医学センターの松田文彦教授、稲富雄一氏、京都大学医学研究科眼科学教室の辻川明孝教授、沼 尚吾助教、大石明生助教（研究当時）、日本 IBM 株式会社の高野敦司氏（研究当時）、中野宏毅氏、壁谷佳典氏、関西医科大学附属生命医学研究所・ゲノム解析部門の日笠幸一郎教授の多大なるご協力を頂きました。ここに深く感謝いたします。

参　考　文　献

[1] Daiger SP. *et al*., Data services and software for identifying genes and mutations causing retinal degeneration, *Invest. Ophthalmol.Vis. Sci.*, 39:S295, 1998.

[2] Oishi M. *et al*., Comprehensive Molecular Diagnosis of a Large Cohort of Japanese Retinitis

Pigmentosa and Usher Syndrome Patients by Next-Generation Sequencing, *Invest. Ophthalmol.Vis. Sci.*, 55(11), pp.7369-7375, 2014.

[3] Abadi M. *et al*., TensorFlow: A System for Large-Scale Machine Learning, In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp.265-283, 2016.

[4] Chollet F., Keras, GitHub, https://github.com/fchollet/keras, 2015.

[5] Kulis B. *et al*., Metric learning: A survey, *Foundations and Trends® in Machine Learning*, 5(4), pp.287-364, 2013.

[6] Kabeya Y. *et al*., Deep convolutional neural network-based algorithm for muscle biopsy diagnosis, *Lab Invest*., 102(3), pp.220-226, 2022.