4J4-1

マイクロアレイデータからの疾患に関連する遺伝子の抽出手法の提案

Finding a Disease-Related Gene from Microarray Data

西脇 一尊 Kazutaka Nishiwaki
金盛 克俊 Katsutoshi Kanamori
大和田 勇人 Hayato Ohwada

東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Tokyo University of Science

Numerous databases on DNA-microarray are now widely available on the internet. Recently, there has been increasing interest in the analysis of microarray data using machine learning techniques due to the number of data that is too massive for researcher to analyze using conventional techniques. In this study, we propose a method of finding a disease-related gene from microarray data using Random Forest. More specifically, we focused on the Alzheimer's disease (AD) and used microarray data related to AD in the experiments. In the result, we found some genes which has been investigated on its relevance to AD, and it proves that our proposed methodology is successful in finding the disease-related gene using the microarray data. In addition, the proposed methodology is useful in providing new knowledge for biologist and medical scientist since there are no previous work on genes that focused on finding a disease-related gene of the Alzheimer's disease.

1. 序論

DNA マイクロアレイ等の遺伝子解析技術の進歩によって，近年遺伝学分野に関する研究は盛んに行われている．多くの実験データはインターネット上で公開され，誰でも利用することができるが，ヒトが持つ遺伝子の数が数万に及ぶと言われている中，それらのデータは人手による解析が難しいほど大きなものとなっている．その中から疾患関連遺伝子の候補を探るのは難しいため，近年この問題に対して機械学習を用いた疾患と遺伝子との関連性の抽出が注目されており，成果を挙げている．例えば，Le Quéau らはアルツハイマー病に関連したマイクロアレイデータからクラスタリングとアソシエーション分析を用いて，発現量の大きい遺伝子間の関連性を見出した[Le Quéau 2014]．しかしこの研究の問題点として，これらの遺伝子同士の関連がアルツハイマー病と関連しているかは不明のままであるという点が挙げられる．また，発現量の小さい遺伝子や，極端に大きい遺伝子間の関連性は見出されていないという点も挙げられる．

そこで本研究では，アルツハイマー病に関連のある遺伝子を抽出することを目的とし，その手法を提案する．本研究では，機械学習アルゴリズムの一つであるランダムフォレスト[Breiman 2001]を用いて，アルツハイマー病患者に関するマイクロアレイデータから疾患との関連が期待される遺伝子の抽出を行う．ランダムフォレストには学習に用いた特徴の重要度を出力することができる．本研究ではこの機能に着目し，健常細胞と非健常細胞を分類する識別器を遺伝子発現量の情報を用いて生成したとき，重要度の高い遺伝子を疾患との関連が期待される遺伝子として抽出を行う．また，マイクロアレイデータに含まれる発現量を正規化することで，発現量の増減が小さい遺伝子についても考慮した候補遺伝子の抽出を行う．

2. 提案手法

本研究における提案手法では，DNA マイクロアレイから得られた各遺伝子の発現量データを用いる．マイクロアレイデータには，各遺伝子の発現量がサンプルごとに行列として記録されており，各サンプルは細胞の健康状態や，疾患の進行度などによって分類されている．図 1 にマイクロアレイデータの構造を示す．

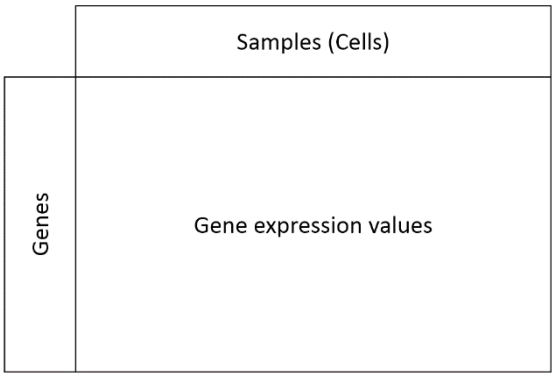

図 1. マイクロアレイデータの構造

ランダムフォレストによる学習を行う際，マイクロアレイデータに含まれる遺伝子の数が数万に及ぶのに対して，サンプルの数は数十程度と非常に少なく，ランダムフォレストが正しく学習を行わない可能性がある．そこで本手法では複数のマイクロアレイデータを 1 つに結合して利用する．その際，結合されるマイクロアレイデータはいずれも同じ身体の部位から得た細胞（脳細胞等）をサンプルとして扱っているマイクロアレイデータのみを用いる．

本手法は，

1. 正規化
2. マイクロアレイデータの結合
3. 事前抽出
4. 疾患に関連のある候補遺伝子の抽出

の 4 ステップから成る．以降の節において，各ステップの詳細を述べる．

2.1 正規化

各マイクロアレイデータによって，実験環境の違いなどから発現量の大きさなどにばらつきが生じているため，発現量データの正規化を行う．ただし，ランダムフォレストによる学習を行う際，学習に用いる特徴量（遺伝子）を厳選するためにサンプル間に

連絡先：西脇 一尊，東京理科大学理工学部経営工学科，千葉県野田市山崎 2641，e-mail：7412105@ed.tus.ac.jp

おける発現量の分散の情報を用いるため，分散の情報を残しつつ正規化を行う必要がある．

サンプルiにおける遺伝子jの発現量をg_{ij}とする．このとき，遺伝子jの発現量ベクトルG_jは以下のようになる．

$$G_j = (g_{1j}, g_{2j}, \cdots, g_{ij}, \cdots)$$

この発現量ベクトルに自身のノルムの逆数を乗ずることで，発現量データの正規化を行う．

$$\frac{G_j}{|G_j|}$$

2.2 マイクロアレイデータの結合

マイクロアレイデータに含まれる遺伝子の数が数万に及ぶのに対して，サンプルの数は非常に少なく，機械学習を行う上で正しく学習が行われない可能性がある．この問題を解決するために，図 2 に示すように各マイクロアレイデータを 1 つに結合することで，学習に用いることのできるサンプルの数を増やす操作を行う．この時，収録されている遺伝子の数はマイクロアレイデータによって異なる場合がある．そこで，各マイクロアレイデータが共通して含む遺伝子の発現量データのみを抽出して結合操作を行い，新たに生成されるデータセットの特徴量として用いる．

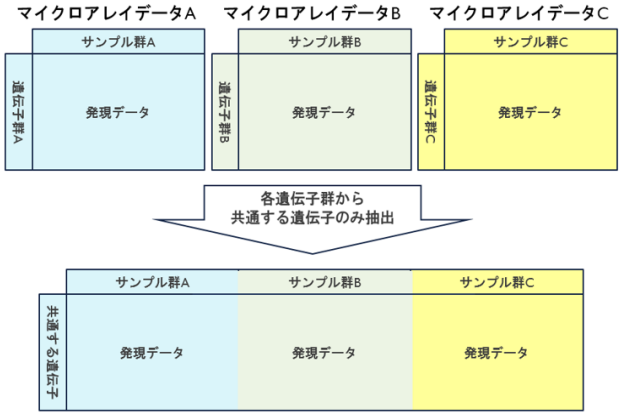

図 2. 複数のマイクロアレイデータの結合

2.3 分散に着目した遺伝子の事前抽出

マイクロアレイデータの結合操作を行った際，各マイクロアレイデータに共通して含まれている遺伝子のみを特徴としてデータセットに取り入れたが，それでもなおサンプル数に対して特徴量の数が大きい．そこで，サンプル間における発現量の分散が大きい遺伝子のみを厳選して学習に用いる．

n個のサンプルがある時，遺伝子αのサンプルごとにおける発現量$\alpha_1, \alpha_2, \cdots, \alpha_n$の平均値をmとした時，分散値

$$V_a = \frac{1}{n}\{(\alpha_1 - m)^2 + (\alpha_2 - m)^2 + \cdots + (\alpha_n - m)^2\}$$

をデータセットに含まれるすべての遺伝子について求め，その値が大きいものから n 個の発現量データを抽出し，ランダムフォレストによる学習の際に用いる．

2.4 疾患関連遺伝子候補の抽出

事前抽出された遺伝子の発現量データを特徴量，サンプルの状態（健常・非健常）をラベルとし，ランダムフォレストによる疾患関連遺伝子候補の抽出を行う．ランダムフォレストの学習はパラメータ設定によって大きく変化してしまうため，データセットごとに最適なパラメータを更新する必要がある．そこで，本手法ではグリッドサーチを用いることで実行毎に最適なパラメータを動的に取得する．本手法では

- 決定木の最大の深さ
- 各ノードに当てはまるサンプル数の最小値
- 葉に当てはまるサンプル数の最小値

をグリッドサーチによって変化させるパラメータとする．表 1 に各パラメータの候補値を示す．

表 1. パラメータ候補値

パラメータ項目	候補値
決定木の最大の深さ	2, 3, 4
各ノードに当てはまるサンプルの最小数	5, 10, 15, 20
葉に当てはまるサンプルの最小数	5, 10

学習を終えたランダムフォレストは，各特徴がどの程度重要なのかを示す重要度を出力することができる．本手法ではこの機能を用いて，重要度の高い遺伝子を疾患との関連が期待される遺伝子として抽出する．

しかし，ランダムフォレストは決定木を生成する際に用いるデータを無作為にサンプリングするため，実行毎に異なる重要度を出力する．そこで，本手法では学習を 5 回繰り返して各遺伝子の重要度を記録し，5 つの重要度の平均値が高い遺伝子から順位付けを行う．

3. 実験

本研究では，インターネット上に公開されているアルツハイマー病についてのマイクロアレイデータを用い，アルツハイマー病との関連が期待される遺伝子を抽出した．本章ではこの実験について述べる．

3.1 データセット

GEO DataSets (http://www.ncbi.nlm.nih.gov/gds/) にて”Alzheimer”をキーワードとして検索し，ヒトを対象生物としていてかつ脳細胞のサンプルを含むマイクロアレイデータのみをフィルタし，ダウンロードした．結果，GDS810，GDS2795，GDS4135，GDS4136，GDS4758 の 5 つのマイクロアレイデータを取得した．

3.2 前処理

各マイクロアレイデータに含まれる共通の遺伝子を抽出する操作を行うが，遺伝子を判別するための ID 番号はデータセットによって異なっていたため，GEO DataSets から各マイクロアレイデータに対するアノテーション情報を参照し，実験由来の ID 番号から共通の Gene ID へ置換を行った．

また，各マイクロアレイデータに同じ遺伝子の発現データが複数記録されている場合があり，遺伝子の数が合わずに結合操作が行えない．そこで，発現量データが複数ある遺伝子について，サンプルごとの発現量の平均値を求め，その値を発現量とする前処理を行った．図 3 に操作の例を示す．

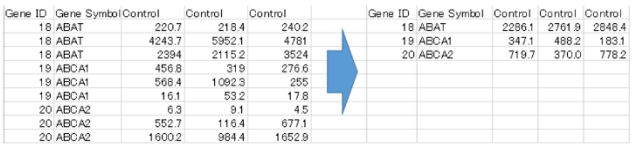

図 3. 重複する発現データの統合操作の例

3.3 マイクロアレイデータの結合と候補遺伝子の抽出

前処理を行った 5 つのマイクロアレイデータについて，発現量の正規化を行った後に結合操作を行った．Gene ID から共通

して含まれている遺伝子を判別し，該当した遺伝子の発現量データを特徴量として 1 つのデータセットを結合した．結果，このデータセットには遺伝子が 11555 件，サンプル数が 178 件（うち非健常サンプル 98，健常サンプル 80）が含まれていた．

このデータセットを用いて，ランダムフォレストを用いてアルツハイマー病に関連する候補遺伝子のランク付けを行った．

4. 結果と考察

実験では事前抽出された 178 の遺伝子について平均重要度を求めた．得られた結果のうち平均重要度が高い遺伝子について，それらがアルツハイマー病との関連が期待されているならば，それについての医学・生物分野の学術論文が存在していると考え，PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) において抽出された遺伝子の名前と，”Alzheimer’s disease”の 2 つのキーワードで OR 検索を行い，2016 年 3 月 7 日時点で登録されている論文の件数を調査した．表 2 に平均重要度が上位 20 件の遺伝子と，PubMed に登録されていた論文の件数を示す．

表 2. 平均重要度が上位 20 件の遺伝子と，アルツハイマー病との関連についての論文件数

Gene ID	遺伝子名	平均重要度	論文件数
2696	GIPR	0.030	0
2068	ERCC2	0.023	3
51599	LSR	0.022	1
26548	ITGB1BP2	0.019	0
8287	USP9Y	0.018	0
83449	PMFBP1	0.018	0
58511	DNASE2B	0.017	0
652	BMP4	0.017	5
56979	PRDM9	0.017	0
6870	TACR3	0.016	1
1828	DSG1	0.015	0
51557	LGSN	0.014	0
8284	KDM5D	0.012	0
26239	LCE2B	0.012	0
3024	HIST1H1A	0.012	0
3375	IAPP	0.012	189
3315	HSPB1	0.011	20
55065	SLC52A1	0.011	0
4057	LTF	0.011	4
147	ADRA1B	0.010	27

その結果，20 種類中 8 の遺伝子について PubMed に論文が登録されていた．特に IAPP（ID:3375）については 189 件と非常に多くの論文が登録されていた．これは，アルツハイマー病の原因として考えられている主な 3 つの仮説[Bartus 1982][Hardy 2002][Schmitz 2004]のうち，アミロイド仮説に関連して IAPP が注目されていることが原因である[Janelle 2014]．従って，本手法が疾患との関連が期待されている遺伝子を抽出することができる事を示した．

一方，実験結果より，GIPR（ID:2696）をはじめ，20 種類中 12 の遺伝子については PubMed に論文が登録されていなかった．これは，現在ではあまりアルツハイマー病との関連は大きくない遺伝子であると考えられるが，本手法により健常・非健常を分類する上では重要な遺伝子であるという事が示され，未知のアルツハイマー病関連遺伝子である可能性がある．以上より，本手法が疾患関連遺伝子の研究において，新たな知見をもたらす可能性があると言える．

ただし，発現量が増加したために罹患したのか，罹患したために発現量が増加したのか，その因果関係は本手法からは見出すことができないという点に留意が必要である．よって，本手法から抽出された遺伝子がそのどちらの因果関係に該当するのかは遺伝学的・疾病学的な検証が別途必要であり，今後の課題として挙げられる．

5. 結論

本研究では，アルツハイマー病に関連のある遺伝子を抽出することを目的とし，遺伝子発現情報を含むマイクロアレイデータからランダムフォレストを用いて，疾患との関連遺伝子の候補を抽出する手法を提案した．

実験では，5 つのアルツハイマー病患者の脳細胞に関するマイクロアレイデータを用いた．5 つのマイクロアレイデータを 1 つのデータセットへ結合し，健常・非健常な細胞を判別する識別器をランダムフォレストにより構築して，そこから出力される各特徴の重要度からアルツハイマー病との関わりが期待される遺伝子を抽出した．

結果，抽出された遺伝子の中にはアルツハイマー病との関わりがあると考えられている遺伝子が含まれており，本手法により疾患関連遺伝子を抽出できることを示した．また，本手法により疾患関連遺伝子の研究において，新たな知見をもたらす可能性を示した．ただし，抽出結果に含まれる遺伝子がアルツハイマー病を引き起こすものなのか，アルツハイマー病によって発現量が増加したものなのか，その判定には遺伝学的・疾病学的な検証が別途必要であり，今後の課題として挙げられる．

参考文献

[Le Quéau 2014] Benoit Le Quéau, Omair Shafiq and Reda Alhajj: Analyzing Alzheimer's disease gene expression dataset using clustering and association rule mining, *Information Reuse and Integration (IRI)* 2014 IEEE 15th International Conference on, Redwood City, CA: pp.283-290.

[Breiman 2001] Leo Breiman: Random forests, *Machine Learning* 45.1(2001): pp.5-32.

[Bartus 1982] Bartus, R. T., Dean, R. L., Beer, B., & Lippa, A. S. (1982). The cholinergic hypothesis of geriatric memory dysfunction. *Science*, 217(4558): pp.408-414.

[Hardy 2002] Hardy, J., & Selkoe, D. J. (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, 297(5580): pp.353-356.

[Schmitz 2004] Schmitz, Christoph, et al. Hippocampal neuron loss exceeds amyloid plaque load in a transgenic mouse model of Alzheimer's disease. *The American journal of pathology* 164.4 (2004): pp.1495-1502.

[Janelle 2014] Janelle N. Fawver et al. Islet amyloid polypeptide (IAPP) a second amyloid in Alzheimer's disease, *Current Alzheimer Research* 11.10(2014): pp.928-940.