4J4-4

機械学習を用いた急性肺障害と関連のあるノンコーディング RNA の抽出

Finding non-cording RNAs related acute lung injury using machine learning

井手 宏武[*1] Hiromu Ide
金盛 克俊[*1] Katsutoshi Kanamori
大和田 勇人[*1] Hayato Ohwada

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

The RNAs are coded in to protein, but there are not being coded in to it. These RNA are called non-coding RNA (ncRNA) and are not become clear it's roles. In this paper, we are proposing a method to find ncRNAs related acute lung injury using microarray data including RNA and ncRNA. In this data, the number of samples is much less than features, so we select some good features, then set 3 groups by elapsed time after administration of a reagent and use Random Forest for each groups. The results suggest that our method predicts at 81.82%, 100.00%, 96.77% accuracy and 91.67%, 100.00%, 100.00% recall and 78.57%, 100.00%, 94.44% precision each groups. The features used in random forest (known RNAs) are not described about acute lung injury in Pubmed, but we may find ncRNAs related the disease.

1. 序論

近年，大規模遺伝子解析技術が急速に進歩し，RNA の網羅的解析がさかんに行われている [Syoji 2003]．しかし，実験によって得た大量のデータから有用なデータを見つけ出すのは難しいため，機械学習によるアプローチが注目されている[Jiang 2007]．ノンコーディング RNA（ncRNA）とは，遺伝子情報をコードしない RNA のことであり，疾患との関わりの研究も進んでいる [Xie 2015]．ncRNA の機能が判明すれば，さまざまな疾患に有効な新薬の開発の促進が期待される．

本研究では，急性肺障害と関連のある ncRNA を抽出することを目的とし，脳で起こす炎症に伴う遺伝子発現変動を記録した mRNA と ncRNA を含むマイクロアレイデータを用いる．本研究では，罹患マウスと非罹患マウスの発現量の値に変動のある遺伝子（変動遺伝子）のみを抽出してからランダムフォレストを実行する．また，抽出した変動遺伝子をランダムフォレストに用いる際に，試薬投与後の経過時間を考慮した 3 グループに分け，各グループにおける変動遺伝子の重要度を比較し，考察する．

2. データ

本研究では，急性肺障害時に脳で起こる炎症に伴う遺伝子発現変動を記録した mRNA と ncRNA を含むマイクロアレイデータを利用する．このデータは，生理食塩水を投与した非罹患マウスと急性肺障害を誘発するためにオレイン酸を投与した罹患マウスの各遺伝子の発現量が，試薬投与後の一定経過時間毎に記録されている．

3. 提案手法

マイクロアレイデータは，サンプル数に比べて遺伝子の数が非常に多いため，ランダムフォレストに用いる遺伝子を選択する必要がある．まず，マイクロアレイデータの遺伝子発現量を正規化し，罹患マウスと非罹患マウスの遺伝子発現量の値に，変動のある遺伝子のみを抽出する．変動遺伝子とする特徴量の評価指標[Umezawa 2013]は(1)式で示す．

$$Z_i = \left| \frac{\bar{x_i} - \bar{y_i}}{\sigma_{x_i} + \sigma_{y_i}} \right| \quad (1)$$

本研究では，全ての遺伝子の(1)式の値の分布から，変動遺伝子の数がサンプル数以下になるような閾値を設定する．さらに，変動遺伝子の核酸配列が空のものを取り除き，残ったものをランダムフォレストの学習データとする．また，試薬投与後の経過時間を考慮し，学習データを 3 つのグループに分ける．group1 は経過時間が比較的早いグループ，group2 は経過時間が比較的遅いグループ，group3 は経過時間を考慮せず，サンプルを全て含めたグループとする．各グループに対してランダムフォレストを実行し,group1,group2 共に重要度が高い遺伝子を疾患と関連のある遺伝子とする．

4. 実験

本実験では，変動遺伝子を抽出してから，試薬投与後の経過時間を考慮してグループ分けし，各グループの遺伝子に対して，罹患マウスか，非罹患マウスであるかの分類予測をランダムフォレストで実行する．また，グループ別に特徴量の重要度も算出する．本実験における各グループの具体的な組み合わせを表 1 に示す．

表 1 経過時間によるグループ分け

	経過時間
group1	1h, 1.5h, 3h, 4h
group2	18h, 24h
group3	1h, 1.5h, 3h, 4h, 18h, 24h

また，ランダムフォレストを実行する際に作成する木の数を 5,000 とし，サンプル数をマウスの数，説明変数は変動遺伝子とする．また，正事例はオレイン酸を投与した罹患マウス，負事例

は生理食塩水を投与した非罹患マウスとする. ランダムフォレスト実行時に各説明変数の重要度も算出し, 重要度の計算にはジニ係数の減少量を利用する.

5. 実験結果

(1)式を用いて変動遺伝子を抽出した結果, その閾値は 0.7 となり, ランダムフォレストに用いる説明変数の数は 12 となった.

表 2 はグループ別のランダムフォレストの結果を示す. サンプル数はランダムフォレストを行う際に用いたサンプルの数を示す. 分割数は交差検定する際に設定したデータの分割数を示す.

表 2 ランダムフォレストの結果

	group1	group2	group3
精度	0.8182	1.0000	0.9677
再現率	0.9167	1.0000	1.0000
適合率	0.7857	1.0000	0.9444
分割数	5	2	7

表 3 は, group1 と group2 における特徴量の重要度の大きさを 3 段階に分け, 比較したものである. 各グループにおいて, 変動遺伝子の重要度が第 1 四分位数以下のものを「低」, 第 1 四分位数より大きく, 第 3 四分位数未満のものを「中」, 第 3 四分位数以上のものを「高」とした. ncRNA は名称が定められていないため, マイクロアレイデータ上の ID で表記した.

表 3 group1 と group2 における変動遺伝子の重要度の大きさの違い

group2 \ group1	高	中	低
高	ID:35073		ID:42746
中		ID:20300 Amfr ID:50831 Tmprss7 Spink12	ID:5946 ID:60892
低	Olfr1329 Dnajb7	Olfr1506	

6. 考察

表 3 より, 変動遺伝子を抽出してから, 試薬投与後の経過時間を考慮してランダムフォレストを行うことで, 精度の高い結果を得られた. また, 経過時間を考慮しない場合も高い精度を示した.

group1 は利用データの中で, 試薬投与後の経過時間が比較的短いサンプルの集合であるが, 3 つのグループと比較して最も精度が低かった. これは, ランダムフォレストに用いた学習データは, 発現量に変動のある遺伝子のみを扱ったので, 試薬投与後の早い時間帯ではすべての遺伝子にある程度の発現量の変動があるので, 多少の誤分類が含まれたと考えられる. 逆に, 試薬を投与してから十分に時間の経ったサンプルの集合である group2 の精度は最も高くなった. これは, 試薬を投与してから十分に時間が経ったので, 遺伝子の発現量に, 分類を行うのに十分な変動があり, 高い精度を示したと考えられる. また, group1 と比較してサンプル数が少なかったことも影響していると考えられる. また, 本実験において変動遺伝子として抽出された mRNA を, 医学系雑誌に掲載された記事や論文を調べることができるデータベースである Pubmed で検索したところ, 急性肺障害との関連性は確認できなかったが, 疾患と関連する遺伝子について新たな知見を得た可能性はあると考えられる.

表 3 から ID:35073 の ncRNA の重要度が group1, group2 共に高い値を示している. また, ID:42746 の ncRNA と Olfr1329, Dnajb7 は試薬投与後の経過時間により重要度に差が出ていることが分かる. 特に, Olfr1329 と Dnajb7 は group1 においては重要度が高く, group2 においては重要度が低くなっている. この 2 つの遺伝子について, 発現にかかる時間が早ければ, 疾患にかかわる遺伝子とは言えないが, group2 において重要度が高い値を示している遺伝子の発現を引き起こす発現である可能性はあると考えられる.

7. 結論

本研究から, 説明変数の数に対してサンプル数が少ないデータにおいて, 罹患マウスと非罹患マウスの分類を高精度で予測することが可能であることを示した. また, 各グループにおいて分類における説明変数の重要度を算出し, 疾患と関連のある遺伝子の候補を示した. 今後の課題として, 本研究で急性肺障害と関連があるとした ID:35073 の ncRNA と疾患の関連性を裏付ける検証実験が必要である.

参考文献

[Syoji 2003] Syoji,Mariko, Mogi,Shinichi: RNA 研究の動向, 科学技術動向,022,pp.9-14, 科学技術政策研究所 科学技術動向研究センター, 2003

[Jiang 2007] Jiang, Peng, et al: classification of real and pseudo microRNA precursors using random forest prediction model with combined features , Nucleic acids research 35.suppl 2, W339-W344, 2007

[Xie 2015] Xie Na, Gang Liu: ncRNA-regulated immune response and its role in inflammatory lung diseases, American Journal of Physiology-Lung Cellular and Molecular Physiology 309.10, L1076-L1087, 2015

[umezawa 2013] Masakazu Umezawa, Keisuke Sekita, et al: Effect of aerosol particles generated by ultrasonic humidifiers on the lung in mouse, Particle and Fibre Toxicology, 2013.