G-021

モチーフ抽出を利用したエピジェネティクス関連領域の予測
Prediction of epigenetic region using motif extraction

東原　正智†
Masanori Higashihara

1. はじめに

エピジェネティクスとは、DAN-RNA-タンパク質という遺伝子の発現以外に化学的な作用で遺伝子の発現を制御する機構である。本研究では、ゲノムワイドに実験された酵母菌の実験データ(遺伝子配列)を用いて、化学的な制御を表すメチル化された部位の特定のため、機械学習を用いて判別分析を行った。

筆者らの研究では、配列の長さを固定長(3 塩基、４塩基)で配列から特徴ベクトルを作成した。しかし、実際にメチル化されている部位の長さは未定である。そのため本研究では、モチーフ抽出を行い、抽出された塩基を特徴ベクトルの作成に用いた。

さらに、特定の塩基の１部位のみが遺伝子発現するということではなく、位置的に離れた塩基の部位が関与して遺伝子発現することが分かっているため、塩基間の関係をも考慮するために上の mRMR の指標を用いた。評価指標として Gini index と属性間の相互情報量を計算する指標である mRMR(Minimum Redundancy and Maximum Relevance)を用いて属性選択を行い、SVM とオンライン学習器である Confidence Weighted learning という２つの学習器を用いた。先行研究では、配列に対して固定長の塩基による sliding window による頻度ベクトルを作成しこれを入力とする研究が多い。しかし、このほかに位置的な情報も判別に寄与すると考えられるために位置を考慮した特徴ベクトルを作成しこれを入力とし、頻度データと位置データでの比較を行った

2. 先行研究

機械学習によるエピジェネティクス関連領域の予測の先行研究としては、Pham らによる SVM を用いた研究がある[11]。彼らは RBF カーネルを用いて予測を行う一方で、別途 polynomial kernel で学習した際の重みを用いて特徴のランキングを行うことにより、特徴ベクトルの属性の重要性を解析している。特徴ベクトルの重要性から属性選択の手法が考えられるがそれについての詳細は[1]に詳しい。多変量解析の主成分分析と SVM を組合せ、属性選択を行いマイクロアレイ解析から癌の判別をする研究[12]も行われている。また、先のような研究では、しばしば高次元データになるため属性選択に関する研究も活発に行われている。新島ら[13]は、化合物とタンパク質の相互作用や活性の予測ばかりではなく、それらに関与する属性を抽出する数理的手法としてカーネル空間での化合物・タンパク質の活性空間を表現し、その空間で特徴抽出する手法を提案している。

†法政大学理工学部経営システム工学科
Email:masanori.higashihara.83@s-adm.hosei.ac.jp

3. 手法

3.1 特徴ベクトル(頻度ベクトルと位置ベクトル)

本研究では、長さ 500 のクロマチンデータを次のように、その位置でカウントした。

(1)特徴ベクトル 1:
モチーフ抽出した塩基をその位置でカウント
(2)特徴ベクトル 2:
3 つまたは 4 つの塩基の組合せ｛AAA,AAT,…,CCC｝{AAAA, …,CCCC｝の 64 種類と 256 種類の組合せをそれぞれの位置でカウント
(3)特徴ベクトル 3:
Gini 係数、mRMR で ranking した上位からの組合せを位置でカウント
(例)AAA,TTT…が上位であれば{AAA}or{TTT}で位置をカウント

特徴ベクトル 2 は、3 塩基の先頭の位置で 1 とカウントしたため最後の 499,500 の位置は 0 となる。特徴ベクトル 3 では、上位の 3 塩基の組合せを用いた。頻度ベクトルについては既発表の長さ 3 または 4 塩基の特徴ベクトルと[11]での長さ 3～6 までの特徴ベクトルと比較した。

図 1 で具体的な特徴ベクトルの例を示す。

（配列データ）

ATCTTTATCTAT…….ATCGGGGAG

(位置)

123456789……500

(特徴ベクトル 1) (CTTT の位置でカウントした場合)

(0,0,1,0,0,0,1,0,0,0,……1,0)

(特徴ベクトル 2) (ATC の位置でカウントした場合)

(1,0,0,0,0,0,1,0,.....0,0)

(特徴ベクトル 3)

(ATC または CTT の位置でカウントした場合)

(1,0,1,0,0,0,1,0,…….……..0,0,0)

図 1.特徴ベクトルの作成

3.2 モチーフ抽出

モチーフ抽出プログラムとして Weeder [9],WordSpy[10]を用いた。前者はベンチマークテストで安定した精度をしめしており、後者は短い配列に対して精度が高い。

3.3 評価指標(Gini index, mRMR)

属性選択は、Gini index では、埋め込み法で randomForest という学習器によって計算されるのでこれを使用した。mPMR については、フィルタ法となるが公開されているプログラムを使用して計算した。以下定義である。

(1) Gini index

Gini 係数は、多様性を示す指標として用いられる。既発表の論文では、random Forest という学習器の中計算される Gini 係数を利用した。

(2) mRMR 最小冗長最大関連指標[2],[7]

(Minimum Redundancy and Maximum Relevance)

定義 1

最小冗長度(Minimize Redundancy)

Min W1, W1=(1/|S|^2) Σ I(i,j)

S:属性数

I(i,j):属性 i と属性 j の相互情報量

最大関連度(Maximum Relevance)

Max V1, V1=(1/|S|) Σ I(h,j)

h: class

I(h,j):クラス h と属性 j の相互情報量

定義 2

MID(相互情報差 Mutual information difference)

Max[I(i,h)-(1/|s|) Σ I(I,j)]

MIQ(相互情報商 Mutual information quotient)

Max[I(i,h)/(1/|s|) Σ I(I,j)]

3.4 機械学習(confidence weighted learning)

先行研究では、学習の予測精度、速度ともに SVM を上回る計算機実験の結果があり、今回 SVM とともに計算機実験に用いた[3], [8]。

4. 計算機実験

実験の手順としては

1. 配列データをモチーフ抽出プログラムに抽出する。
2. 機械学習による判別分析を行う。
3. 属性選択の指標として Gini 係数、mPMR を計算しフィルタ法による選択を行う。
4. 属性選択の指標による順位から上位の属性を組合せ特徴ベクトルを作成し、実験を行う。

OS	Ubuntu Linux 8.04
CPU	Intel Xeon 2.50GHz(Core2 Duo)
Memory	4GB
HDD	500GB×2

表 1.計算機環境

データセットとしては、Pokholok のデータ[6]から実験を行い高い精度をしめしたデータ[14]を用いた。

データ	説明	正例	負例
H3	H3 ヒストンの存在確率	7667	7298
H4	H4 ヒストンの存在確率	6480	8121

表 2.データセット

5. まとめ

既発表の論文[4],[5]では、頻度を用いた特徴ベクトルと位置情報を用いた特徴ベクトルを作成し比較をおこなった。その結果、判別分析の予測率に関しては頻度ベクトルが位置情報を用いた特徴ベクトルより高い。しかしこれらは固定長であり、今回のモチーフ抽出を用いることにより高い精度の同定ができると考えられる。また、位置的に離れた塩基間が起こす反応についても今回 mRMR を利用することにより一歩近づいていると考えている。属性の評価指標によるランキング、予測率など実験結果の詳細については発表当日に行う。

参考文献

[1] Yvan Saeys,Inaki Inza and Pedro Larranaga, “A review of feature selection techniques in bioinformatics”, Bioinformatics, Vol.23, No.19 (2007).

[2] Hanchuan Peng,Fuhui Long,and Chris Ding “Feature Selection Based Information: Criteria of Max-Dependency, Max-Relevence.and Min-Redundancy”, IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol.27, No.8 (2005).

[3] Mark Dredze, Koby Crammer, and Fernando Pereira, “Confidence-Weighted Linear Classification”, ICML, (2008).

[4] Higashihara,M.,Rebolledo-Mendez,J.D.,Yamada,Y.,Satou,K. Application of a Feature Selection Method t o Nucleosome Data: Accuracy Improvement and Comparison with Other Methods, WSEAS Transactions on Biology and Biomedicine, Vol.5, Issue 5, pp.95-104,2008.5.

[5] 東原正智,位置情報を用いたエピジェネティクス関連領域の予測と属性選択,DEIM2010.

[6] D.K. Pokholok et al., Genome-wide Map of Nucleosome Acetylation and Methylation, *Cell*, Vol.122, pp.517-527.

[7] http://penglab.janelia.org/proj/mRMR/index.htm

[8] http://code.google.com/p/oll/wiki/OllMainJa

[9] http://159.149.109.9/modtools/

[10] Guandong Wang, Taotao Yu and Weixiong Zhang, WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar, Nucleic Acids Research,Vol. 33,2005.

[11] T.H. Pham, D.H. Tran, T.B. Ho, K. Satou and G. Valiente, Qualitatively Predicting Acetylation and Methylation Areas in DNA sequences, *Genome Informatics*, Vol.16, No.2, 2005, pp.3-11.

[12] Fabian Model, Péter Adorján , Alexander Olek et al,Feature selection for DNA methylation based cancer classification,Bioinformatics Vol. 17 no. 90001 2001,Pages p.157-p.164.

[13] 新島 聡, 奥野 恭史, 化合物-タンパク質活性空間における特徴選択,IBIS2009.

[14] http://www.jaist.ac.jp/~tran/nucleosome/