G-007

頻出系列パターンを用いた遺伝子発現予測の予備的検討
A preliminary study on explanatory gene expression models with frequent sequential pattern mining

中野 玄[1)]　山本 泰生[1)]　守屋 央朗[2)]
Gen Nakano　Yoshitaka Yamamoto　Hisao Moriya

1　研究背景と目的

1.1　研究背景

生命には DNA と呼ばれるアデニン (A)，チミン (T)，シトシン (C)，グアニン (G) の 4 つの塩基からなる塩基配列が存在する．この塩基配列上に存在する生命を形作る情報を遺伝子と呼ぶ．DNA からタンパク質を生成する一連のプロセスはセントラルドグマと呼ばれ，複雑多様な生命現象を説明する最も重要な基本原理とみなされる．生命科学分野において，モデル単細胞生物からヒトゲノムを含め数多くの DNA が解読されてきた．しかし，解読した各遺伝子型がどのような表現型を持つのかという問いに対しての理解は部分的なものにとどまっている．この課題を本稿では遺伝子-表現型課題と呼ぶ．

文献 [1] では，大規模変異株探索 (DMS) と呼ばれる実験法と Transformer に基づく深層学習モデルを用いて遺伝子-表現型課題に取り組んでいる．DMS とは，ランダムな塩基置換を導入した DNA を「遺伝子型入力」，増殖速度や遺伝子発現量などを「表現型出力」として，これらの入力-出力の関係性を大規模に調査する実験手法である．DMS によって網羅的な大規模データを取得することができる．Vaishnav ら [1] はこの大規模データを用いて，Transformer ベースの遺伝子発現予測モデルを学習させることで，高精度な遺伝子発現予測を実現できることを示した．

一方で，深層学習モデルの問題として，遺伝子-表現型課題を理解するために不可欠な解釈性の欠如が挙げられる．

1.2　目的

本研究では，解釈可能性を持つ系列パターンを用いて遺伝子-表現型課題に取り組む．はじめに，系列パターンに基づく発現量予測モデルを構築し，その性能評価を行う．次に，複数の照合方法による系列パターンの頻度分布の調査を行い，識別能力の高い系列パターンの網羅的に探索する．上記より，発現量を有意に決定する塩基列が存在するかどうか検証する．

2　パターン抽出

2.1　データ

文献 [1] で収集された塩基配列データを用いる．塩基配列データは，プロモータ領域上の 80 文字の塩基配列とその塩基配列を持つ変異株における蛍光タンパク質 (YFP) の遺伝子発現量のペアで構成される．データ数は約 3000 万である．今回の問題設定は二値分類とし，遺伝子発現量が最小区分 (0 番ビン) にあるデータと最大区分 (17 番ビン) にあるデータを対象としてパターン抽出，予測，頻度分布の調査を行う．

1）　静岡大学大学院 総合科学技術研究科
2）　岡山大学学術研究院 環境生命自然科学学域

2.2　データの前処理

図 1 のように，系列長 80 の塩基配列データに対し起点順序分解 [2] を行い，位置情報を付加した塩基配列データを得る．図中の例では，1 番目の塩基 A が 1A というアイテムに変換されている．このように位置情報を付加することで系列パターンを位置情報付きのアイテムの集合に変換し，これにより効率的な頻出アイテム集合マイニング法を用いて系列パターンを抽出する．

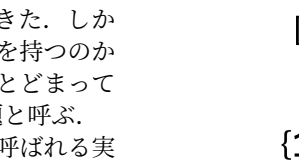

図 1　起点順序分解により位置情報を付加した塩基配列の集合表現

2.3　パターン抽出手法

2.3.1　CICLAD

CICLAD[3] はスライディングウィンドウ型の頻出アイテム集合マイニング法である．与えられた入力データに対して走査範囲のウィンドウをスライドさせていき，パターンと頻度情報を厳密に抽出する．頻度情報とは，パターンがそのウィンドウ中に何回出現したかという情報を示す．CICLAD は厳密なマイニング法であり，抽出されるパターン数が膨大にとなる．このため，プロモータ領域の塩基配列データを 20 塩基ごとに 3 分割しパターン抽出を行う．分割したデータをそれぞれ 20L(左部)，20C(中央部)，20R(右部) と呼称する．

2.3.2　PARASOL

PARASOL[4] はランドマークウィンドウ型の頻出アイテム集合マイニング法である．対象となるデータ全体に対して，上位 k 件の頻出なパターンを近似的に求める．PARASOL は近似マイニング法でありデータサイズによらず高速にパターンを抽出することができる．本研究では 20 塩基からなる部分塩基配列と全塩基配列のそれぞれのデータで頻出パターンを抽出する．ただし，PARASOL の出力は近似解であり，抽出パターンの頻度として実測値と推定値の 2 つが利用できる．マイニングにおいて対象のパターンが実際に出現した回数が実測値であり，対象のパターンが抽出対象のデータ全体で出現すると推定される回数が推定値である．

3　パターンを用いた予測

3.1　予測手法

3.1.1　最長系列に基づく予測

既存研究 [2] では，最長系列に基づく予測手法が提案されている．パターンおよびパターンの頻度を用いて予測を行う．すなわち，指定した最低頻度以上の頻度をもつパターンのうち，パターン長が最大のものを一つ選択

し，そのパターンの抽出元データの発現量 (0 番ビンもしくは 17 番ビン) を予測とする．また，パターン照合において各塩基の位置情報を考慮する場合としない場合の 2 通りで性能評価する．

3.1.2 Passive Aggressive アルゴリズムを用いた予測

Passive Aggressive アルゴリズム (PA 法)[5] を用いた予測手法を新たに提案する．PA 法を用いた予測手法では，系列パターン，頻度，遺伝子発現量に予測価値を加えた 4 項組のデータを扱う．データはパターン抽出用データ，テストデータ，予測価値訓練用データの 3 種に分割して扱う．予測価値の更新について以下に手順とアルゴリズムを示す．

訓練データと系列パターンを用いて予測価値の更新を行う．まず，訓練データに出現する系列パターンのうち最長のものを探す (8 行目)．次に，出現した最長系列パターンの予測価値を遺伝子発現ラベルごとに合計し，予測価値の合計が高い方を予測値とする (9 行目)．訓練データの遺伝子発現ラベルの真値と予測値を比較する (10 行目)．真値と予測値が等しかった場合，何もしない．真値と予測値が異なっていた場合，真値と等しい遺伝子発現ラベルを持つ最長系列パターン集合の予測価値を増やす (11, 12 行目)．これを繰り返し，正しい予測が行えるように予測価値を更新する．

訓練データによる更新後の予測価値を用いてテストデータの予測を行う．PA 法によって，高い予測価値を持つ系列パターンが発見できる．それらの系列パターンは遺伝子発現量を決定づける有意な塩基列と仮定できる．

Algorithm 1 PA 法における予測価値の更新

Input: $DNA = \langle [ed_1, l_1^t], \ldots, [ed_i, l_i^t] \rangle$: 塩基配列, 遺伝子発現ラベル真値からなる 2 項組データ
$Patt = \langle [ep_1, ex_1, pv_1], \ldots, [ep_j, ex_j, pv_j] \rangle$: パターン, 遺伝子発現ラベル, 予測価値からなる 3 項組データ
Output: $Patt = \langle [ep_1, ex_1, pv_1], \ldots, [ep_j, ex_j, pv_j] \rangle$: 予測価値更新後のパターンデータ

1: **Function** $MaxPatt(ed, Patt)$
2: 塩基配列に出現するパターンのうち，パターンの長さが最大であるパターンデータ集合を返す関数
3: **End Function**
4: **Function** $UpdateValue(Patt)$
5: パターンデータ集合の予測価値の値を上げる関数
6: **End Function**
7: **for** $m = 1, 2, \ldots, i$ **do**
8: $T = MaxPatt(ed_m, Patt)$
9: $l^p = Predict(T)$
10: **if** $l_m^t \neq l^p$ **then**
11: $T' = \{[ep, ex, pv] \in T \mid ex = l_m^t\}$
12: $UpdateValue(T')$
13: **end if**
14: **end for**

3.2 性能評価

3.2.1 実験設定

実験の構成を以下の表 1，表 2 に示す．抽出条件と予測条件の番号は紐づいている．抽出条件の入出力数は遺伝子発現量 0 と 17 のデータの合計である．予測条件 1 においては，テストデータのデータセットを 3 種類作成し平均を算出した．予測条件 2，3，4 では，PARASOL によるパターン抽出を行なっているため，頻度は実測値と推定値の 2 種類存在する．

表 1 パターン抽出の各条件

条件	抽出手法	入力数	塩基数	出力数
抽出 1	CICLAD	6000	20	約 2700 万
抽出 2	PARASOL	約 73 万	20	2 万,10 万
抽出 3	PARASOL	約 73 万	80	2 万,10 万
抽出 4	PARASOL	約 58 万	80	2 万,10 万

表 2 予測の各条件

条件	予測手法	パターン	テストデータ
予測 1	最長系列	抽出 1	1500
予測 2	最長系列	抽出 2	約 15000
予測 3	最長系列	抽出 3	約 15000
予測 4	PA	抽出 4	約 15000

3.2.2 実験結果

以下の表 3 は予測条件 1 での実験結果である．表 4 は予測条件 2 での実験結果である．

表 3 位置情報の有無による予測条件 1 の F1 値

	20L	20C	20R
位置情報あり	0.72	0.70	0.70
位置情報なし	0.68	0.70	0.70

表 4 予測条件 2 の F1 値

パターン数 (頻度)	20L	20C	20R
20000(実測値)	0.64	0.62	0.61
20000(推定値)	0.61	0.62	0.60
100000(実測値)	0.66	0.63	0.61
100000(推定値)	0.63	0.60	0.61

以下の表 5 は予測条件 3，4 での実験結果である．

表 5 予測条件 3 と予測条件 4 の F1 値

パターン数 (頻度)	予測条件 3	予測条件 4
20000(実測値)	0.63	0.65
20000(推定値)	0.64	0.65
100000(実測値)	0.65	0.66
100000(推定値)	0.64	0.66

3.3 考察

予測に関して，予測条件 3 と予測条件 4 を比較すると，どの場合でも F1 値が上がっている．したがって，PA 法を用いた予測手法の優位性を示していると考える．予測条件 4 において頻度実数値と頻度推測値の差が少ない点については，予測価値という情報を追加したことによって頻度情報の結果に与える影響が少なくなったものと考える．

4 パターンの照合

4.1 照合の目的

得られたパターンを塩基配列と照合し，パターンの頻度分布を調べる．また，照合方法による統計量の変化も確認する．これらによって，識別能力のあるパターンの発見を目指す．

4.2 照合の定義

ここで行う照合とは，パターンと塩基配列の一致を確認することである．一致とは，対象パターンの各要素が対象の塩基配列の中に出現していることを指す．各パターンで一致した塩基配列の個数を頻度とする．塩基配列全体における各パターンの頻度を求める．3 種の照合方法を用いて照合を行う．

4.3 照合方法

4.3.1 厳密照合

一つ目の方法は厳密照合である．パターンの位置および塩基が塩基配列に出現していた場合，そのパターンが塩基配列と一致していたとみなす．例えば，パターン{3A, 5C}は{1C, 2T, 3A, 4G, 5C}と一致しているが，{1C, 2T, 3G, 4G, 5C}と一致していない．

4.3.2 位置揺らぎ照合

二つ目の方法は位置揺らぎ照合である．許容可能な位置の誤差 W を設定する．パターンの各要素において位置が W の範囲で塩基配列に出現していた場合，そのパターンが塩基配列データと一致していたとみなす．例えば，誤差 W を 1 に設定する．パターン{3A, 5C}は{1C, 2A, 3T, 4G, 5C}と厳密照合では一致していない．一方でパターン{3A}に対応する塩基配列{2A}があるため，位置揺らぎ照合で一致しているとみなす．以降の実験では $W = 1$ と設定する．

4.3.3 スライド照合

三つ目の方法はスライド照合である．パターンの各要素の位置の差に着目した照合方法である．パターンの要素間の位置の差を保ったまま，パターンの位置を 1 から 80 までスライドさせる．そのうちいずれかで位置と塩基が塩基配列に出現していた場合，そのパターンが塩基配列データと一致していたとみなす．例えば，パターン{3A, 5C}は{1C, 2A, 3T, 4C, 5C}と厳密照合で一致はしていない．一方でパターン{3A, 5C}に対応する塩基配列{2A, 4C}があるため，スライド照合で一致しているとみなす．

4.4 実装方法

4.4.1 ビットマップと垂直配置の利用

実装方法を述べる．塩基配列データの情報を図 2 のように垂直に配置し，ビットマップ形式で保持する [6]．すなわち赤枠で囲まれたビット列は 1 番目の塩基が A である塩基配列データの各インデックスを示している．このような垂直配置データを用いて照合を行う．垂直配置を用いることで，論理演算によるパターン照合が行え，演算処理の高速化が見込まれる．

4.4.2 照合方法のアルゴリズム

垂直配置を用いた厳密照合のアルゴリズムを Algorithm2 に示す．

4.5 照合結果

4.5.1 実験設定

使用するデータは遺伝子発現量が 0 番ビンと 17 ビンの 2 種類がある．またパターンマイニング法 (PARASOL) の実行には抽出パターン数の上限 $k = 10000$ と $k = 50000$ の 2 種類がある．よって 2 つの掛け合わせから 4 種類のパターンデータが生成される．照合方法は上記に挙げた 3 種類ある．それぞれの照合方法の結果を以下に示す．

塩基配列	1A	1C	1G	1T	…	80A	80C	80G	80T
1A2C3T…79G80T	1	0	0	0	…	0	0	0	1
⋮	0	1	0	0	…	0	1	0	0
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮
⋮	1	0	0	0	…	1	0	0	0

図 2 垂直配置ビットマップ

Algorithm 2 厳密照合

Input: $DNA = \langle ed_1, \ldots, ed_i \rangle$: 垂直配置・ビットマップ形式で保持した塩基配列データ,
$patt = \langle ep_1, \ldots, ep_j \rangle$: パターンデータ

Output: $result = \langle er_1, \ldots, er_j \rangle$: 各パターンデータで照合が一致した回数 (頻度)

1: **Function** $k(\alpha)$
2: パターンの要素 α に対応する塩基配列 (ビット列) のインデックスを返す関数
3: **End Function**
4: **for** $m = 1, 2, \ldots, j$ **do**
5: $t = \{k(\alpha) | \alpha \in ep_m\}$ (要素に対応するインデックス番号を取得)
6: $TMP := \{ed_u | u \in t\}$ (DNA から ep_m に合致する ed を抽出する)
7: $R := \bigwedge TMP$ (TMP 中の各 ed_u を AND 演算)
8: $er_m := |R|$ (得られたビット列 R のうち，ビットが 1 のものを数える)
9: **end for**

4.5.2 分布図

以下の図 3 は遺伝子発現量 17 および 0 のパターンデータの分布を各頻度 60000 までの範囲で切り出した分布図である．図 4，図 5，図 6 はそれぞれ，照合一致回数が遺伝子発現量 17 と 0 の塩基配列データのどちらかに偏っているものを抽出した分布図である．縦軸は遺伝子発現量 17 の塩基配列における頻度，横軸は遺伝子発現量 0 の塩基配列における頻度である．赤点は遺伝子発現量 17 のパターン，青点は遺伝子発現量 0 のパターンである．赤の破線は傾きの 3 の直線であり，青の破線は傾き 1/3 の直線である．赤 (青) の破線より上 (下) にある赤点 (青点) のみを選択して表示させている．

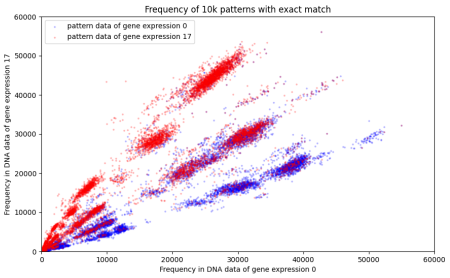

図 3 パターン数 10k 厳密照合 分布比較 切り出し

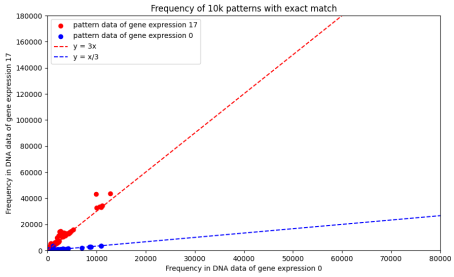

図 4　パターン数 10k 厳密照合

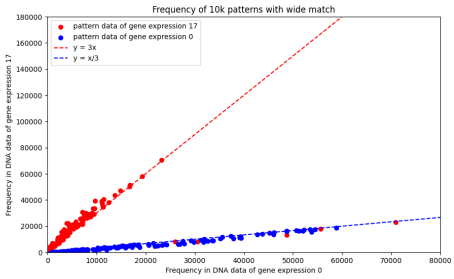

図 5　パターン数 10k 位置揺らぎ照合

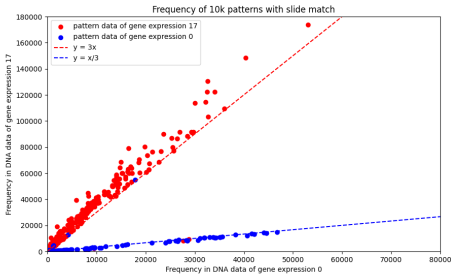

図 6　パターン数 10k スライド照合

4.5.3　統計量

以下の表 6 は $k = 10000$ かつ遺伝子発現量 17 における各パターンと照合方法の統計量である．ここでの比率とは，遺伝子発現量 17 の塩基配列での出現頻度と遺伝子発現量 0 の塩基配列での出現頻度の割合を示す．比率は遺伝子発現量 0 の塩基配列データにおける頻度が最低 1 以上のパターンの値に絞って算出した．スライド照合に関して，パターンの要素数が 1 の場合全塩基配列と一致してしまう．そのため，パターンの要素数 2 以上のものを対象として頻度と比率を算出した．

表 6　パターン数 10k 遺伝子発現量 17 の統計量

	比率 3 以上のパターン数	最大の比率	全体の平均頻度	全体の比率平均
厳密	693	19	21457	1.7
位置揺らぎ	108	9.5	111496	1.31
スライド	441	34	249281	1.38

4.5.4　高速化結果

垂直配置・ビットマップによる高速化の結果を示す．以下の表 7 に示すように，約 35 倍高速化した．

表 7　垂直配置・ビットマップによる高速化結果

	平均時間 (秒)
線形探索 (厳密照合)	99.8
垂直配置 (厳密照合)	2.84

4.6　考察

表 3 から，遺伝子発現量 0 と 17 のパターンデータでは，分布が異なることが読み取れる．このことから，パターンによる予測で示されていたパターンデータ全体の識別能力が分布の差という形で改めて示された．

表 6 の統計量から位置揺らぎ照合とスライド照合の特徴が読み取れる．位置揺らぎ照合では，比率 3 以上のパターン数，最大の比率共に減っており，条件を緩めたことで遺伝子発現量 0 および 17 の塩基配列どちらにも一致した結果が現れている．一方のスライド照合では比率 3 以上のパターン数をあまり減らさずに最大の比率の値を上げている．また，スライド照合で全体の平均頻度が高い理由は，要素数の少ないパターンでの一致率が高かったものだと考える．

5　まとめと今後の展望

本研究では，系列パターンによる遺伝子発現予測法の性能評価と系列パターンの頻度分布調査を行った．予測においては，PA 法を用いた提案手法による精度向上という結果を得た．照合においては，各照合方法の頻度分布と統計量を得た．また，高速な実装を行なった．

今後の展望として，PA 法により発見された予測価値の高い系列パターンや，照合によって得られた出現の偏りが顕著な系列パターンの識別能力の調査を行いたい．また，既知の DNA 配列モチーフとの比較を行い，系列パターンと生物学的意味の結合可能性を探る追加実験を行いたい．

参考文献

[1] Vaishnav, Eeshit Dhaval, et al., The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, vol. 603, no. 7901, pp. 455–463, 2022.

[2] Hinano Sako and Yoshitaka Yamamoto, Online Closed Episode Mining with Root-Order Decomposition, in *2023 IEEE International Conference on Big Data (BigData)*, pp. 2683–2689, 2023.

[3] Martin, Tomas et al., CICLAD: A fast and memory-efficient closed itemset miner for streams, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1810–1818, 2020.

[4] Yamamoto, Yoshitaka et al., PARASOL: a hybrid approximation approach for scalable frequent itemset mining in streaming data, *Journal of Intelligent Information Systems*, vol. 55, pp. 119–147, 2020.

[5] Crammer, Koby et al., Online Passive-Aggressive Algorithms, *Journal of Machine Learning Research* vol. 7, no. 19, pp. 551–585, 2006.

[6] 宇野毅明, 有村博紀, AI レクチャー 「頻出パターン発見アルゴリズム入門-アイテム集合からグラフまで-」人工知能学会全国大会論文集, pp. 413–413, 2008.