東京大学大学院新領域創成科学研究科
情報生命科学専攻

平成 18 年度

修士論文

植物ゲノムのプロモーター領域の特徴解析と予測

2007 年 3 月提出
指導教員　高木　利久　教授

56911　巻島　健志

概要

遺伝子の転写制御は、発生や分化などのすべての生命現象の根幹をなすものである。転写制御の仕組みを知るためには、まず各種の転写因子がプロモーター領域のどのような特徴を認識しているのかを明らかにしていかなければならず、プロモーター領域の特徴解析は大きな課題の一つである。近年、シロイヌナズナとイネといった植物において大量の完全長 cDNA データが利用可能になったことにより、転写開始点（TSS）を正確に決定でき、プロモーター領域のゲノムワイドな解析が可能になった。本研究では、主にシロイヌナズナとイネにおける種間及び種内でのプロモーター領域の塩基配列の共通性と多様性を明らかにするため、特徴解析とプロモーター領域の予測を行った。

特徴解析においては、繰り返し配列、GC 含量、GC-skew、CpG/CpNpG アイランド、エントロピーという特徴に注目し解析を行った。これまで、植物のプロモーター領域（TSS 周辺）には２塩基から６塩基の短い繰り返し配列が多数存在すること事が報告されていた。本研究では、繰り返し配列のないプロモーター領域でも、di-mer や tri-mer といった配列が高頻度に存在し、これらは TSS 周辺の GC-skew や GC 含量の推移と大きく相関があることを明らかにした。このような領域は転写因子によって認識されている可能性がある。CpG/CpNpG アイランドはこれまでシロイヌナズナのプロモーター領域に存在するとされてきたが、これらはイネにも存在し、プロモーター領域よりもむしろ UTR や遺伝子領域の N 末端周辺に多数存在することが分かった。また、イネのプロモーター領域はシロイヌナズナのプロモーター領域に比べてエントロピーが高く、塩基配列の多様性をもつことを明らかにした。

次に、特徴解析の結果を踏まえ、２次のマルコフモデルと重回帰分析による位置のスコアに対する重み付けの併用によって、コアプロモーター領域の予測を行った。これは単純な手法であるにもかかわらず、シロイヌナズナにおいては感度 88.66%、特異度 89.63%という高い予測精度を示した。このことは、プロモーター領域周辺の塩基組成が種内で共通していることを示していると考えられる。イネにおいては、tri-mer の出現頻度に基づく k-means クラスタリングをさらに使用することで、感度 84.63%、特異度 76.55%という予測精度を示した。各クラスタはそれぞれ異なる特徴をもち、全体のうち約 20%を占める一つのクラスはシロイヌナズナと類似した特徴をもつことを明らかにした。今後、他の植物のデータが利用可能になれば、進化的に保存されたプロモーター領域を明らかにすることも可能になるだろう。

目次

第１章

序論

真核生物における発生及び分化過程は、遺伝子の発現（転写）の段階で多くの制御が行われている。RNA ポリメラーゼⅡによって転写が開始されるためには、転写開始点（TSS）の近くに存在するコアプロモーターを転写基本因子が認識し、転写開始複合体が形成される必要がある。次に、転写開始複合体がコアプロモーターの上流に存在する転写制御領域を認識して転写活性化因子が結合し、転写装置が形成されて転写が開始される（図 1.1）。これら転写制御の仕組みを知るためには、まず各種の転写因子がプロモーター領域のどのような特徴を認識しているのかを明らかにしていかなければならず、プロモーター領域の特徴解析は大きな課題の一つである。

真核生物において、転写基本因子によって認識されるコアプロモーターには TATAbox や Initiator といったモチーフが存在することが知られている。しかしながら、TATAbox をもつコアプロモーターは多くても全体の約半数に過ぎず(Shahmuradov *et al* 2003、Carlos and Erich 2005)、Initiator に関しても数塩基程度のモチーフに過ぎない。

そのため、まずはプロモーター領域周辺に存在するその他の特徴を明らかにする必要がある。さらに、その特徴を明らかにするにはプロモーター領域を同定する必要がある。近年、オリゴキャップ法（Suzuki *et al* 1997）、キャップラッパー法(Carninci *et al* 1996)というような mRNA の両端をより正確に解析する技術が開発されることでその問題が大きく解決された。この技術はヒト（Yudate *et al* 2001、Kikuno *et al* 2002）、マウス（Okazaki *et al* 2002）、ショウジョウバエ（Rubin *et al* 2000）といった生物にて適応され、完全長の cDNA の配列を得られるようになった。この cDNA 配列をゲノム配列にマップすることで、正確かつ大量に転写開始点（TSS）を決定でき、プロモーター領域のゲノムワイドな解析が可能になった。この結果、哺乳類では CpG アイランドというプロモーター領域を強く示唆する特徴が発見され、この特徴によってプロモーター領域の予測プログラムの精度が大きく向上した(Davuluri *et al* 2002、Down and

Hubbard 2002、Bajic *et al* 2003、Bajic *et al* 2004)。

その後、シロイヌナズナ（Seki *et al* 2002)、イネ（Kikuchi *et al* 2003）といった植物で完全長の cDNA のデータが使えるようになった。それに伴い、プロモーター領域のゲノムワイドな解析が行われ、シロイヌナズナとイネを使った解析により繰り返し配列が TSS 周辺に存在すること(Fujimori *et al* 2003)、シロイヌナズナを使った解析により CpG アイランドがプロモーター領域に多数存在すること（Rombauts *et al* 2003)、シロイヌナズナを使った解析によりコアプロモーター領域にいくつかのモチーフが存在すること(Carlos and Erich 2005)などの解析が行われている。

しかしながら、植物のプロモーター領域のゲノムワイドな解析はまだ始まったばかりであり、更なるプロモーター領域の特徴解析を行う必要がある。

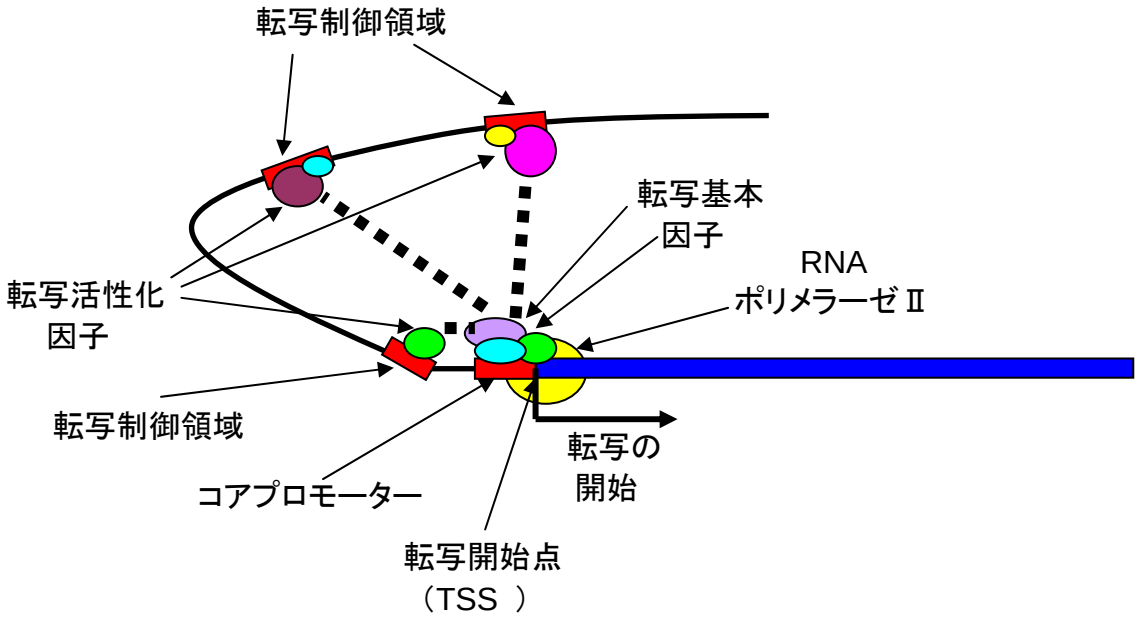

図 1.1 遺伝子調節領域の略図

第２章

プロモーター領域の特徴解析

2.1 背景と目的

植物は、多様な発生様式を持ち、その結果できる形態も多様である。発生過程の進化は転写因子の多様化によって説明できるのではないかと考えられており、実際、シロイヌナズナの解析によって 3000 個以上の転写に関わる遺伝子が発見され、そのうち半分以上が転写因子であるという報告がされている(Seki *et al* 2002)。プロモーター領域と転写因子の関係を明らかにするためには、まずはこれら多様な転写因子が認識しうるプロモーター領域の特徴を知らなければならない。

現在、完全長の cDNA の解析によって、mRNA の両端が正確に決定され、これらの配列を使って転写開始点（TSS）がより正確にかつ大量に決定されるようになった。このことによって、コアプロモーターを含む TSS の上流下流約数百 bp の領域について解析が行われるようになりつつある(Carlos and Erich 2005)。

植物では２塩基から６塩基の短い繰り返し配列（マイクロサテライト）が TSS から 5'UTR の領域にかけて多数存在することが報告されている(Fujimori *et al* 2003)。このような短い繰り返し配列は、哺乳類では TSS や 5'UTR よりイントロンの領域に多数存在する。しかしながら、この研究では繰り返し配列の種類（繰り返す塩基の種類や個数など）ごとに詳細な解析が行われていない。

ゲノム上において、CpG というシトシン（C）とグアニン（G）が連続した配列が高密度に存在する領域を CpG アイランド(Island)といい、遺伝子発現の調節、インプリンティング等に関係していると言われている（Jeddeloh *et al* 1998、Kooter *et al* 1999、Bender 2001、Richards *et al* 2002）。CpG という配列では C がメチル化されていること多いのだが、CpG アイランドの C はあまりメチル化されておらず、哺乳類においてはプロモーター領域周辺に多数存在することが分かっている（Feltus *et al* 2003）。植

物では CpG だけでなく、CpNpG、その他の３塩基（CpHpH、H=A,T,C）といった領域であるにおいても C のメチル化が起こることが知られている（Antequera *et al* 1988、Finnegan and Kovac 2000）。近年、CpG アイランドと CpNpG アイランドがシロイヌナズナのプロモーター領域を示唆できるという報告がある（Rombauts *et al* 2003）。しかしながら、これらがイネなどその他の植物について当てはまるかどうかについては検証を行う必要がある。

この章ではコアプロモーター領域を含む TSS から上流 1000bp から下流の 500bp までの領域を主な解析対称とする。これらの領域にはコアプロモーターはもちろんのこと、その上流に存在する転写制御領域、TSS 下流から開始コドンまでの領域である 5'UTR の領域、遺伝子領域のＮ末端周辺の領域などを含んでいる。植物におけるプロモーター領域の特徴を明らかにするために、先に述べた繰り返し配列の種類の違いによる詳細な解析やイネのプロモーター領域における CpG/CpNpG アイランドの有無をはじめとし、GC 含量、GC-skew やエントロピーといった特徴にも注目して解析を行った。

2.2 方法

2.2.1 データセット及び特徴解析領域の抽出法

シロイヌナズナとイネの 2 種類の植物を主な解析対象にした。これは先に述べたようにシロイヌナズナとイネの 2 種で完全長の cDNA の解析が進んでいることに加えて完全長のゲノム配列が決定しているためである（Arabidopsis Genome Initiative 2000、Goff *et al* 2002）。

シロイヌナズナについては、cDNA の配列を RARGE(http://rarge.gsc.riken.jp/)から、ゲノム配列を NCBI からそれぞれダウンロードして利用した。RARGE は完全長の cDNA の解析によって得られた配列と EST 配列から cDNA の 5'側（TSS）を決定し、その cDNA 配列と TSS から上流 100bp までの領域を格納しているデータベースである。1 遺伝子座につき一つの cDNA 配列が格納されており、冗長性の無いデータである。ゲノム配列は Accession 番号が NC_003070、NC_003071、NC003072、NC003075、NC00306 の配列を利用した。それぞれ染色体の 1 番から 5 番までの配列に相当する。なお、ゲノムサイズは約 120Mbp である。

イネについては、cDNA 配列を TIGR(http://www.tigr.org/)から、ゲノム配列を IRGSP（http://rgp.dna.affrc.go.jp/J/IRGSP)からそれぞれダウンロードして利用した。TIGR は完全長の cDNA 配列によって得られた配列によってオルタナティブスプライシングなどを考慮して 1 遺伝子座につき複数の cDNA 配列が格納されている。IRGSP は Build 4.0 を使用した。染色体数は 12 本であり、ゲノムサイズは約 390Mbp である。

先に述べた cDNA 配列を BLAST（Altschul *et al* 1990）、sim4(Florea *et al* 1998)を用いてゲノム配列にマップし、そのゲノム上にマップされた cDNA 配列の 5'側の位置を TSS として決定した。イネの場合は 1 遺伝子座につき複数の cDNA 配列が格納されているため、マップされた 5'側のうち最も頻度が多かったものを TSS とした。決定した TSS の上流 1000bp から 下流 500bp の領域[-1000,+500]をプロモーター領域の正解とし主な解析対称とした。なお、この領域はコアプロモーター、UTR、転写制御領域を含んでいる。作成したデータ数はシロイヌナズナが 15166 個、イネが 18731 個であった。また、プロモーター領域以外のデータセットを作るために[-2000,+1000]の領域をすべてマスクし、それ以外の領域から長さ 1500bp の配列を抽出し不正解セットとした。なお、十分な領域が取れなかったもの、ATGC 以外の塩基を含むものはデータセットから取り除いた。

2.2.2 繰り返し配列

繰り返し配列の抽出には Sputnik(http://espressosoftware.com/pages/sputnik.jsp)を利用した。ここでは、図 2.1 を用いて抽出方法を説明する。例えば 2 塩基の繰り返し配列を抽出したいとして、配列の 1 番目と 2 番目に CT という塩基配列があったとする。3 番目の配列が 1 番目の配列と同じ C ならばスコアに＋1 を与え、それ以外の塩基ならば－6 を与える。次に 4 番目の配列が 2 番目の配列と同じ T ならば＋1 を加え、－6 を与えるという操作を繰り返すことによって繰り返し配列を抽出する。

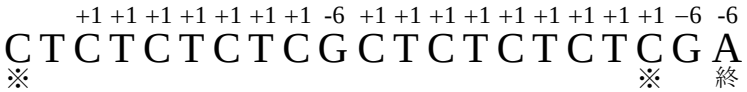

図 2.1

このアルゴリズムにしたがって、すべての領域を探索する。ただし、そのスコアが 0 以下になったら探索を止める。最終的にスコアが 9 以上になった配列のみを抽出し、繰り返し配列とした。図 2.1 の場合では※～※の間でスコアは 10 となり C と T の 2 塩基の繰り返し配列となる。

繰り返し配列として抽出されるためには、2 塩基の繰り返し配列は長さ 11bp 以上、3 塩基の繰り返し配列は長さ 12bp 以上、3 塩基の繰り返し配列は長さ 13bp 以上、5 塩基の繰り返し配列は長さ 14bp 以上が最低限必要である。

2.2.3 GC-skew

GC-skew の値は与えられた塩基配列中の C、G の塩基配列の出現頻度によって、以下の式よって導かれる。$n(N)$は与えられた配列に存在する塩基配列 N の個数を示している。

$$\text{GC - skew} = \frac{n(C) - n(G)}{n(G) + n(C)} \qquad (2.1)$$

塩基配列の頻度を計算するウィンドウのサイズは 100bp で、1bp ずつずらして各領域の値を計算した。

2.2.4 CpG/CpNpG アイランド

ある長さ 200bp の領域の GC 含量が 50%以上で、領域内に存在する CpG が観測値/期待値(Obs/Exp)が 0.6 以上の場合に、その領域を CpG アイランドと定義している（Gardiner-Garden and Frommer 1987)。観測値とは単純に CpG の表れる回数で、期待値とはその時の領域中の C と G の頻度によって導き出される期待値である。同様に CpNpG アイランドも長さ、GC 含量、観測値/期待値によって導くことが出来る。CpNpG アイランドに関してはこれら３つの閾値は定義されていない。

今回はLarsen(1992)らによって開発されたCpGアイランドの抽出プログラムを使用した。このプログラムに改良を加え、CpNpG アイランドも抽出可能にした。なお、このプログラムはCで書かれている。

2.2.5 情報エントロピー（情報量）

シャノンのよって定義された情報量の考えに基づき、与えられた固定長の配列データセットにおいて、地点 i における各領域での k-mer(k=2~5)の出現確率によって以下のように求めることができる。

$$H_i(P_i) = -\sum_{N \in \Omega} P_i(N) \log_2 P_i(N) \qquad (2.3)$$

ここで、$P_i(N)$はある地点 i での k-mer(N)の出現確率である。Ω は k-mer の起こりうるすべて配列で、例えば k=３のときに Ω＝｛AAA,AAC,…,TTG,TTT｝となる。

この指標は直感と一致よく一致するものである。例えば、与えられた配列群の地点 i における 3-mer の出現確率 P_iによってエントロピーを求めた時に、P_i＝1（すべてのデータの地点 iが同一の k-mer である）ならばエントロピーは 0 で特徴的な領域といえる。なお、P_i＝1/64（配列が完全にランダムに並んでいる）ならばエントロピーは 6 である。

以上、特にプログラムの明記が無い場合には、perl、java、R、を使用して解析を行った。

2.3 結果と考察

2.3.1 繰り返し配列

植物の TSS から 5'UTR の領域には繰り返し配列が存在することが知られている(Fujimori *et al* 2003)。しかしながら、繰り返す配列の数や塩基組成や位置といった詳細な解析は行われていない。今回はこれらについて詳細な解析を行った(表 2.1、表 2.2)。

プロモーター領域のデータセット（正解セット）のうち、少なくとも一つ繰り返し配列をもつデータ数はシロイヌナズナで 4470(29.49%)、イネで 5983(31.94%)であった。プロモーター領域以外のデータセット（不正解セット）においてはシロイヌナズナで 2037(13.43%)、イネで 2095(11.18%)であった。不正解セットと比較して、シロイヌナズナでは２塩基と３塩基の繰り返し配列が多く、イネでは３塩基の繰り返し配列が最も多く観察された。このことから、植物のプロモーター領域には２～３塩基程度の短い繰り返し配列が頻出することが再確認された。

表 2.1 TSS 周辺[-100,200]に存在する繰り返し配列(シロイヌナズナ)

	正解セット		不正解セット	
繰り返しの種類	個数	割合(%)	個数	割合(%)
２塩基	1821	12.01	645	4.26
３塩基	2344	15.46	759	5.01
４塩基	342	2.26	272	1.80
５塩基	550	3.63	458	3.02
2∪3	3861	25.46	1372	9.05
2∪3∪4∪5	4470	29.47	2037	13.43

繰り返しの種類について、２～５塩基とは少なくとも一つ２～５塩基の繰り返し配列を持つデータの数と割合を表している。２∪３は２塩基もしくは３塩基の繰り返し配列、２∪３∪４∪５は少なくとも一つ繰り返し配列を持つデータの数と割合と表している。

表 2.2 TSS 周辺[-100,200]に存在する繰り返し配列(イネ)

	正解セット		不正解セット	
繰り返しの種類	個数	割合(%)	個数	割合(%)
２塩基	1380	7.37	625	3.34
３塩基	3935	21.01	893	4.77
４塩基	984	5.26	396	2.12
５塩基	1272	6.80	386	2.07
2∪3	4820	25.73	1460	7.79
2∪3∪4∪5	5983	31.94	2095	11.18

繰り返しの種類について、２～５塩基とは少なくとも一つ２～５塩基の繰り返し配列を持つデータの数と割合を表している。２∪３は２塩基もしくは３塩基の繰り返し配列、２∪３∪４∪５は少なくとも一つ繰り返し配列を持つデータの数と割合と表している。

次に、繰り返し配列の長さと観察される詳細な位置について検証を行った(図 2.2、2.3)。ここでは、TSS から少し離れた転写制御領域などについても繰り返し配列が観察できるか確認するため、領域を[-1000,500]にした。どちらの植物とも正解セット中の 2 塩基の繰り返し配列は TSS 周辺に多いのに対して、3 塩基の繰り返し配列は TSS の下流にかけて多いことが分かる。特に、イネの 3 塩基の繰り返し配列は TSS の下流に広範囲に渡って観察された。シロイヌナズナでは５塩基の繰り返し配列が TSS の上流 400bp 周辺で若干観察され、イネでは３塩基の繰り返し配列は TSS の上流数百 bp に渡って観察された。これらの領域の繰り返し配列は転写制御領域の特徴を示している可能性がある。

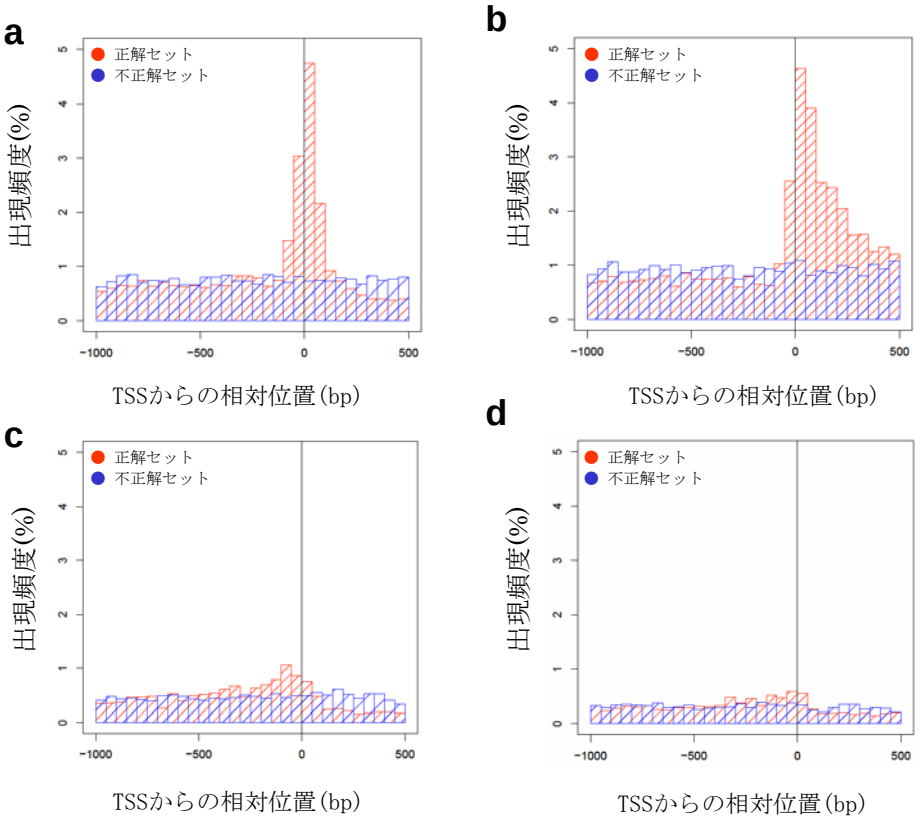

図 2.2 繰り返し配列の出現位置（シロイヌナズナ）。縦軸は繰り返し配列の出現頻度(%)、横軸は TSS からの相対位置(bp)。TSS を中心として各領域で繰り返し配列を検出したデータの割合を示している。**a** ２塩基から成る繰り返し配列。**b** ３塩基から成る繰り返し配列。**c** ４塩基から成る繰り返し配列。**d** ５塩基から成る繰り返し配列。

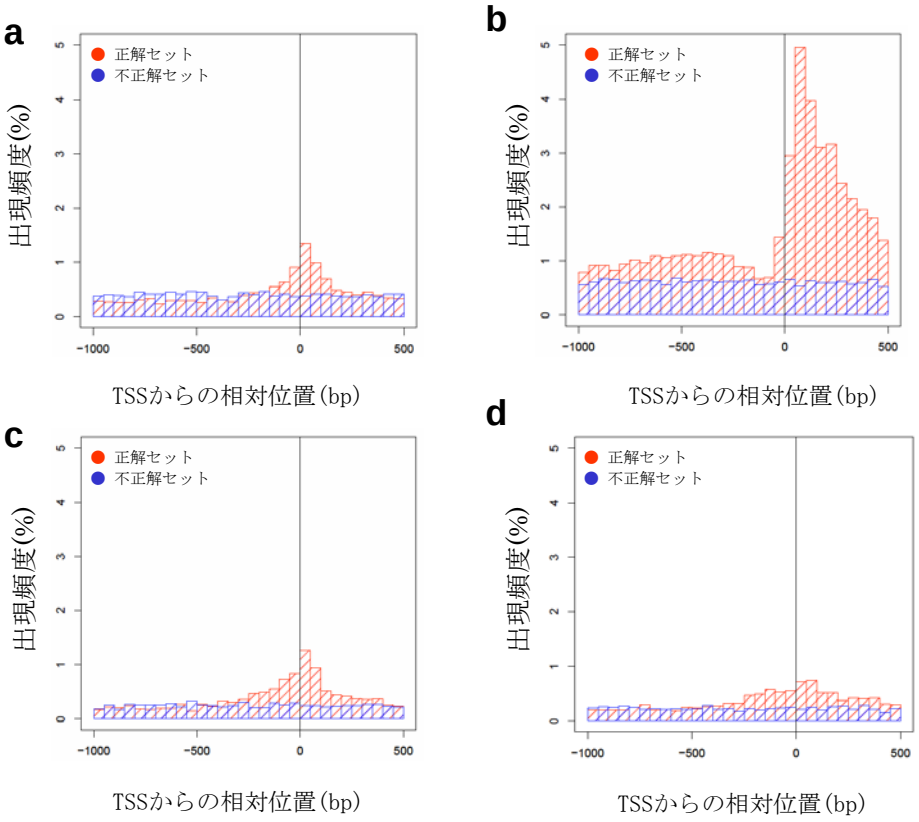

図 2.3 繰り返し配列の出現位置（イネ）。縦軸は繰り返し配列の出現頻度(%)、横軸は TSS からの相対位置(bp)。TSS を中心として各領域で繰り返し配列を検出したデータの割合を示している。**a** ２塩基から成る繰り返し配列。**b** ３塩基から成る繰り返し配列。**c** ４塩基から成る繰り返し配列。**d** 5 塩基から成る繰り返し配列。

プロモーター領域[-1000,500]に繰り返し配列が存在することが確認できた。先の解析では TSS 周辺の領域は繰り返す塩基の長さによってのみ分類し観察を行った。一見すると２種の間で共通性が高いように見えるが、同じ２塩基の繰り返し配列であったとしても構成される塩基組成は異なっている可能性もある。よって、次に個別の塩基の種類の違いによってどのような繰り返し配列が存在するのか検証した（表 2.3、2.4）。

表 2.3 繰り返し配列の種類と個数（シロイヌナズナ）

種類	正解セット中の個数	不正解セット中の個数	正解/不正解
CG	1	0	∞
CT	1223	96	12.74
ACG	42	4	10.50
CGT	33	4	8.25
CTT	1235	179	6.90
CGG	12	2	6.00
CCT	128	24	5.33
AG	476	133	3.58
ATC	182	59	3.08
AGC	24	8	3.00
AAG	463	168	2.76
CCG	17	7	2.43
AAC	140	60	2.33
ACC	44	19	2.32
AC	93	46	2.02
ACT	17	9	1.89
CTG	9	7	1.29
AGG	39	33	1.18
GT	38	33	1.15
GTT	55	61	0.90
ATT	38	46	0.83
ATG	43	53	0.81
AGT	11	15	0.73
GGT	19	30	0.63
AAT	34	56	0.61
AT	152	376	0.40

表 2.4 繰り返し配列の種類と個数（イネ）

種類	正解セット中の個数	不正解セット中の個数	正解/不正解
CCT	712	51	13.96
CTT	172	24	7.17
AAG	90	15	6.00
CCG	1154	200	5.77
ACC	240	43	5.58
AGC	188	41	4.59
CGG	859	200	4.30
CT	554	133	4.17
AGG	247	59	4.19
CGT	149	42	3.55
CTG	138	40	3.45
ACG	98	29	3.38
GGT	105	32	3.28
CG	161	61	2.64
AC	98	40	2.45
ACT	34	14	2.43
AAC	18	8	2.25
GTT	27	14	1.93
AGT	33	18	1.83
AG	271	158	1.72
ATC	35	21	1.67
GT	51	32	1.59
AAT	33	24	1.38
ATG	24	18	1.33
ATT	30	31	0.97
AT	166	192	0.86

ここでは、２塩基と３塩基の繰り返し配列についてのみの結果を示した。青はＣと Ｔ のみから成る繰り返し配列、緑はＣと Ｇのみから成る繰り返し配列、黄色はＡと Ｇのみから成る繰り返し配列、赤はＡと Ｔのみから成る繰り返し配列を示している。

C と T のみから成る繰り返し配列が 2 種で多いことが分かる。また、C と T の相補鎖である A と G の 2 塩基のみから成る繰り返し配列も比較的検出されている。逆に、A と T のみから成る繰り返し配列は 2 種とも少ない。C と G のみから成る繰り返し配列はシロイヌナズナではあまり検出されないにもかかわらず、イネでは多数検出されている。また、(CG)n,(AGC)n,(CCG)n,(CGG)n,(CTG)n といった CpG/CpNpG アイランドに類似した繰り返し配列もイネで多数観察された。以上から、C と T からのみ成る繰り返し配列の存在はシロイヌナズナとイネのプロモーター領域に共通した特徴で、C と G のみから成る繰り返し配列の存在はイネのプロモーター領域において特徴的であることが明らかになった。

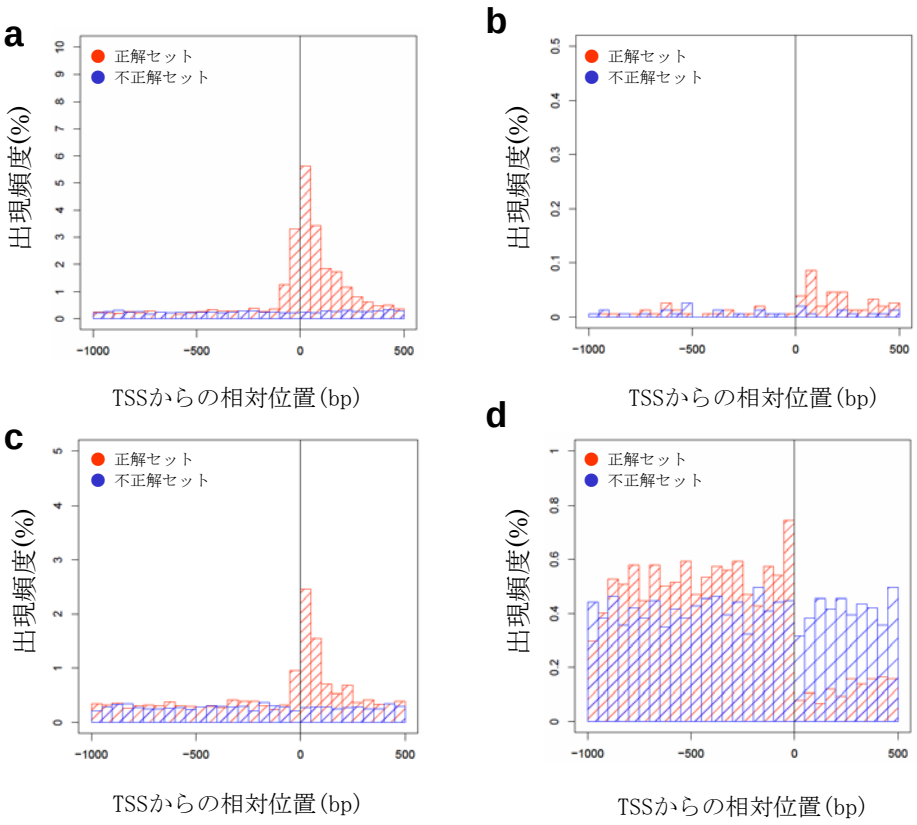

図 2.4 繰り返し配列の出現位置（シロイヌナズナ）。縦軸は繰り返し配列の出現頻度(%)、横軸は TSS からの相対位置(bp)。**a** C と T のみから成る繰り返し配列。**b** C と G のみから成る繰り返し配列。**c** A と G のみから成る繰り返し配列。**d** A と T のみから成る繰り返し配列。

次に、先ほどと同様に繰り返し配列が観察される詳細な位置について検証を行う。C と T のみ、C と G のみ、A と G のみ、A と T のみから成る繰り返し配列が TSS を中心として、どの領域に存在するのか観察した（図 2.4、2.5）。２種の植物で共通して、C と T のみから成る繰り返し配列が TSS 周辺の上流 100bp から TSS の下流にかけて多く観察された。イネで多数存在した C と G のみから成る繰り返し配列は TSS から下流にかけて観察された。

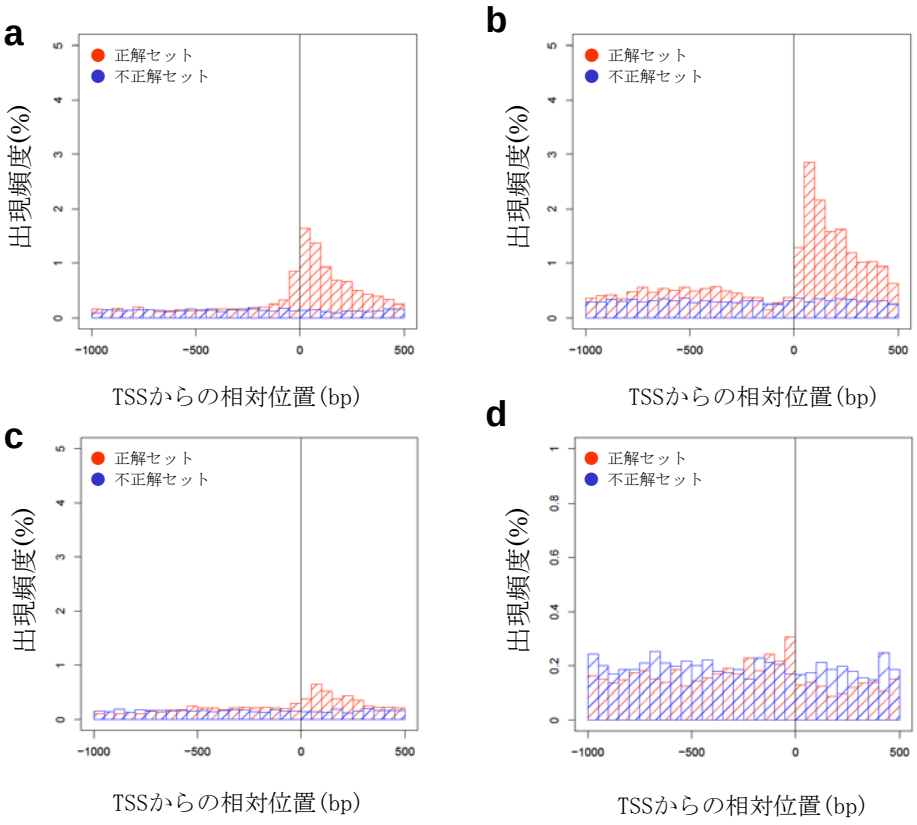

図 2.5 繰り返し配列と TSS からの相対位置（イネ）。縦軸は繰り返し配列の出現頻度(%)、横軸は TSS からの相対位置(bp)。**a** C と T のみから成る繰り返し配列。**b** C と G のみから成る繰り返し配列。**c** A と G のみから成る繰り返し配列。**d** A と T のみから成る繰り返し配列。

以上の結果から、シロイヌナズナとイネでは２種に共通して２~３塩基の比較的単純な繰り返し配列が多いことを確認した。さらに、その繰り返し配列を構成する塩基の種類に偏りがあることが分かった。その中でも、２種に共通して C と T のみから成る繰り返し配列が TSS 周辺に多数存在している。また、種に特異的な例として、シロイヌナズナの TSS 下流では A と T のみから成る繰り返し配列が少ないことや、イネの TSS 下流では C と G のみから成る繰り返し配列が多数存在していることが観察された。

2. 3. 2 GC 含量と繰り返し配列

文献から抽出したプロモーター領域の配列を格納した The Eukaryotic Promoter Database （Schmid *et al* 2004）というデータベースがある。このデータベースには多種の植物のプロモーター領域が格納されており、それらのデータにおいては GC 含量が TSS 上流で減少し、TSS 下流で増加することが確認されている(Kanhere and Bansal 2005)。まず正解セットと不正解セットを用いてこの傾向を確認する（図 2.6）。

２種ともに TSS 下流で GC 含量が増加することが確認された。TSS 上流で GC 含量が減少することはシロイヌナズナでは確認できたが、イネではバックグラウンドをそれほど変わらなかった。ただ、TSS の上流 200bp の周辺では、TSS の他の領域[-600,300]に比べて減少している。イネについては TSS 下流の GC 含量の増加は C と G のみから成る繰り返し配列と同じ領域で観察されている(図 2.5b)ことから、イネの正解セットの繰り返し配列と GC 含量の関係を検証するため、TSS 周辺に繰り返し配列があるものとないものにデータセットを分けて GC 含量の推移を観察した(図 2.7)。シロイヌナズナについても確認のため同様に GC 含量の推移を観察した。

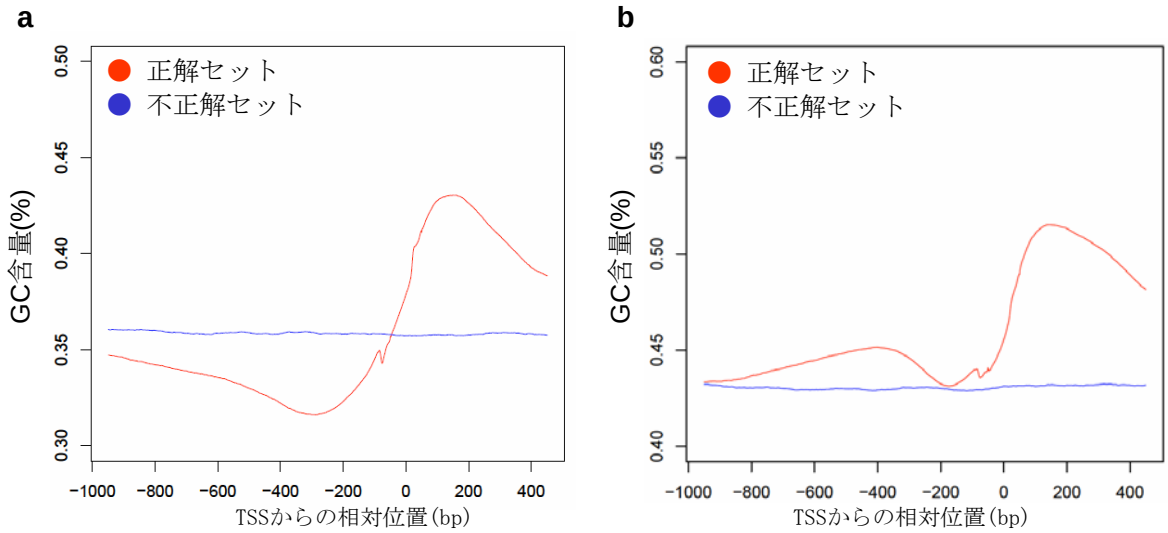

図 2.6 GC 含量の推移。縦軸は GC 含量(%)の値、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

シロイヌナズナについては、C と G のみから成る繰り返し配列がほとんど検出されなかったことからも予想できたように、GC 含量は繰り返し配列の有無でほとんど推移に差が無かった。一方、イネついては、繰り返し配列の有無で異なる推移を観察した。具体的には繰り返し配列のあるデータセットにおいて TSS 下流 200bp 周辺で大きく GC-skew が増加した。また、イネの繰り返し配列のないデータセットにおいて TSS 上流 100bp 周辺で GC 含量が減少し、TSS 上流 400bp 周辺で GC 含量が増加している。

結果、イネの正解セットの GC 含量の推移を次のように説明できる。TSS 上流 400bp 周辺の値の増加と TSS 上流 100bp 周辺の値の減少は繰り返し配列のないデータセットに起因し、TSS 下流 200 周辺の値の上昇は繰り返し配列のあるデータセットに起因している可能性がある。

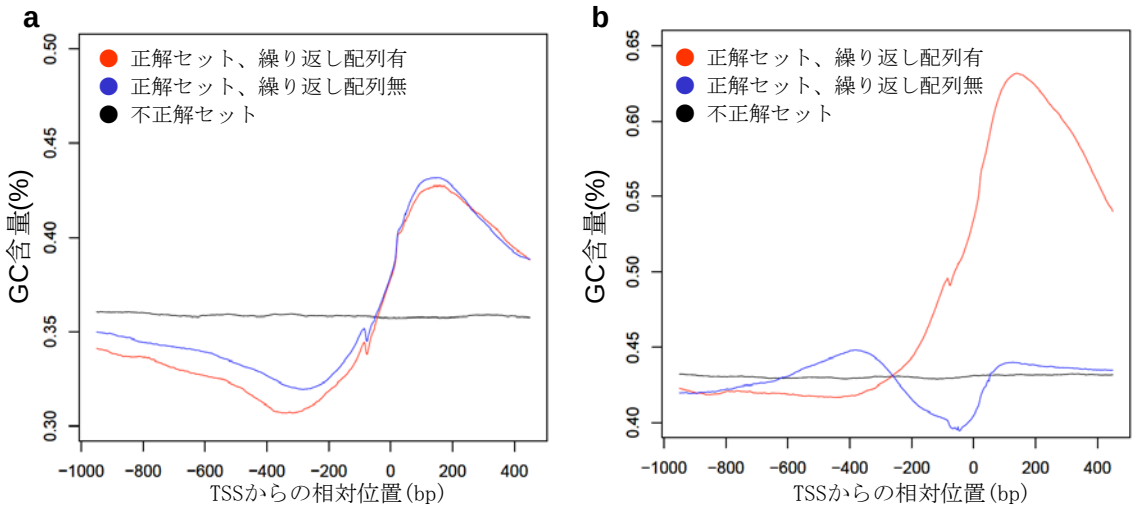

図 2.7 繰り返し配列の有無による GC 含量の推移の違い。縦軸は GC 含量(%)の値、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

2. 3. 3 GC-skew と繰り返し配列

植物のプロモーター領域（TSS 周辺）では、シロイヌナズナとイネについて GC-skew が増加する、すなわち C と G の出現頻度に差があることが確認されている(Fujimori *et al* 2005)。ここでは、まず正解セットと不正解セットを用いてこの傾向を確認する（図 2.8)。２種とも、GC-skew の増加は TSS 周辺で観察された。イネのデータセットにおいて、GC-skew の値が TSS 上流 200bp でバックグラウンドより低くなることを観察した。

次に、この GC-skew の推移が先にその存在を確認した繰り返し配列によるものか否かを検討する。まず、TSS 周辺に繰り返し配列があるものとないものに正解セットを分けて GC-skew の推移を観察した（図 2.9)。シロイヌナズナについては、GC-skew の推移が繰り返し配列の有無でほとんど差が無かった。一方、イネついては、繰り返し配列の有無で異なる推移を観察した。具体的には繰り返し配列のあるデータセットにおいて TSS 周辺で大きく GC-skew が増加した。この推移はシロイヌナズナと類似した推移を示している。また、イネの繰り返し配列のないデータセットにおいて TSS 下流 200bp で GC-skew が減少した。

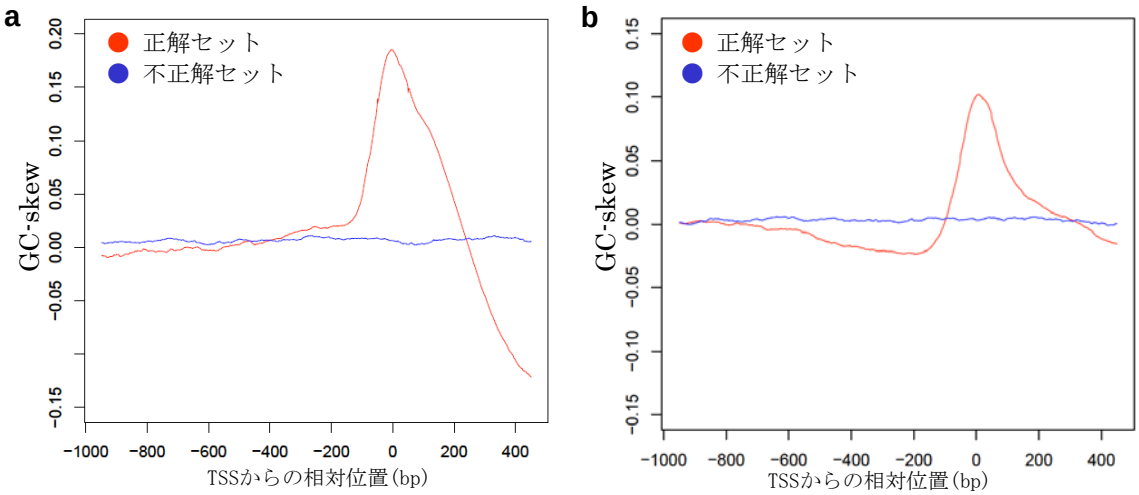

図 2.8 GC-skew の推移。縦軸は GC-skew の値、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。 **b** イネ。

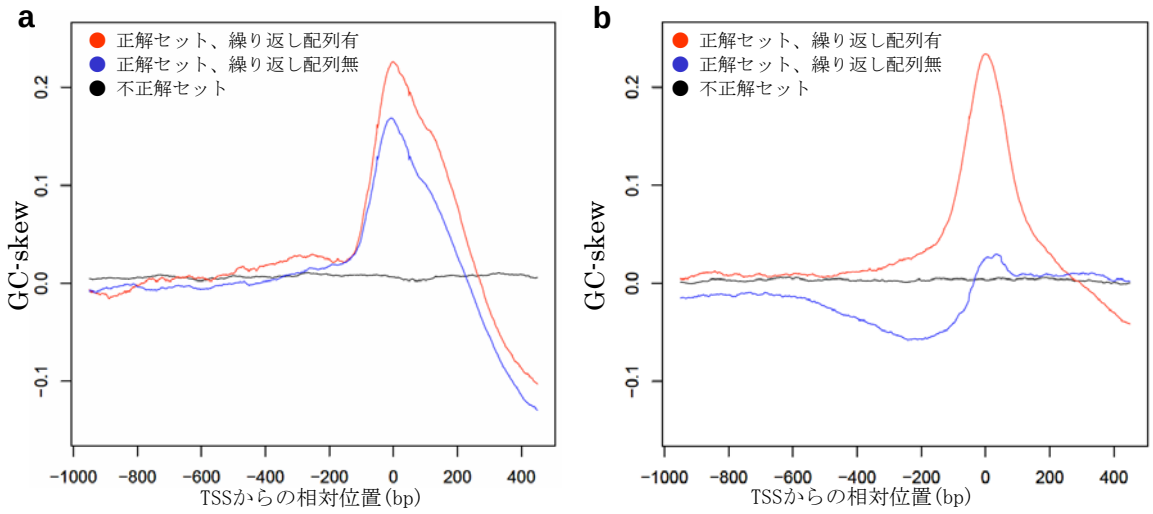

図 2.9 繰り返し配列の有無による GC-skew の推移の違い。縦軸は GC-skew の値、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

結果、イネの正解セットの GC-skew の推移を次のように説明できる。TSS 上流 200bp 周辺の値の減少は繰り返し配列のないデータセットに起因し、TSS 周辺の値の上昇は繰り返し配列のあるデータセットに起因している可能性がある。

シロイヌナズナのプロモーター領域には多数の C と T のみから成る繰り返し配列があることから、繰り返し配列のあるデータセットで GC-skew が増加することは予想できた。しかしながら、シロイヌナズナの繰り返し配列がないデータセットにおいても GC-skew が観察された。この原因を探すため、繰り返し配列の有無でデータセットを分けて di-mer の頻度を観察した（図 2.10）。ここでは C と T のみから成る繰り返し配列を多数確認していたことから CT と TC の頻度を観察したものを示す。

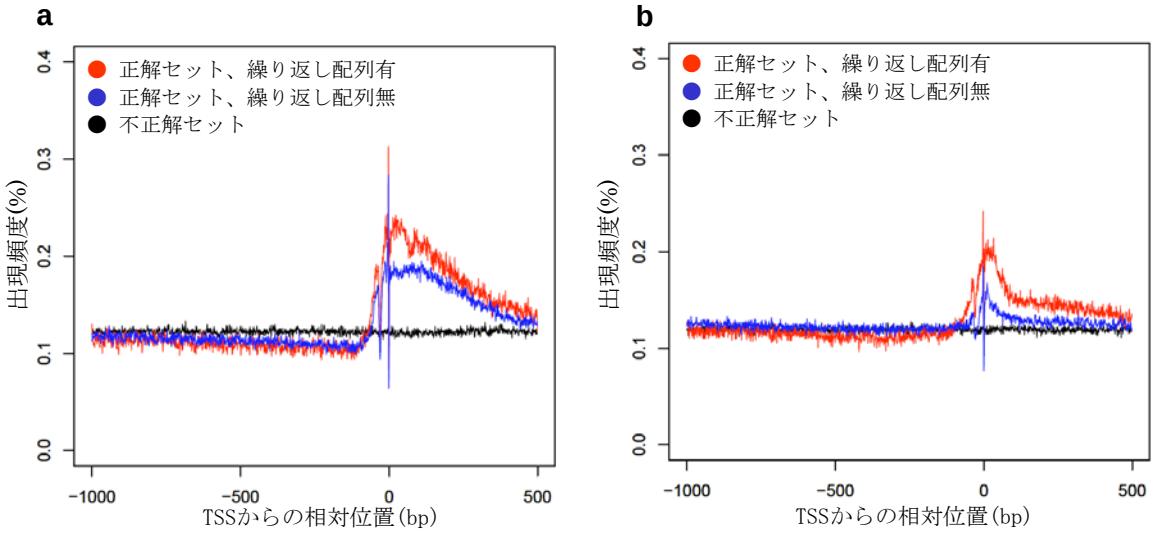

図 2.10 繰り返し配列の有無による CT と TC の出現頻度の違い。縦軸は TSS からの各位置における CT と TC の頻度（%）、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

シロイヌナズナの繰り返し配列のないデータセットにおいても、TSS 周辺に繰り返し配列のあるデータセットと同程度の CT と TC を観察した。このことから、TSS 周辺での GC 含量と GC-skew の変化は di-mer の出現頻度の偏りである可能性が考えられる。さらに、繰り返し配列の有無に関わらず、同様の発現制御配列をもつ可能性が考えられる。一方、イネでも繰り返し配列のないデータセットで TSS 周辺の CT と TC の出現頻度が若干多くなっていた。しかしながら、繰り返し配列の有無で比較すると顕著な差が確認できることから、これらは異なる発現制御配列をもつ可能性が考えられる。

2.3.4 CpG/CpNpG アイランド

シロイヌナズナではプロモーター領域で CpG/CpNpG アイランドがイントロン領域や遺伝子間領域と比べて多数存在することが報告されている（Rombauts *et al* 2003）が、イネでの解析はまだ行われていない。ここでは、イネの GpG/CpNpG アイランドが観察可能かどうか、また、どの領域に観察できるのか検証した。

各生物で塩基組成には大きな違いがあり、生物種ごとに同一の定義を使って CpG/CpNpG アイランドを検出することは出来ない。ここでは先に述べた定義の 3 つの指標（GC 含量が 50%、領域の長さが 200bp、観測値/期待値が 0.6）を基準として、一つの指標のみを変化させながら CpG アイランドを検出した。変化させた幅はそれぞれ GC 含量では 10%、領域の長さ（length）では 100bp、観測値/期待値(Obs/Exp)では 0.3 ずつである。さらに CpNpG アイランドも CpG アイランドと同様の定義を基準とし、GC 含量、領域の長さ、観測値/期待値をそれぞれ変化させて CpNpG アイランドを検出した。CpG アイランドに関して、シロイヌナズナでは GC 含量が 40%、長さが 200bp、観測値/期待値が 1.5、イネでは GC 含量が 70%、長さが 400bp、観測値/期待値が 0.9 の時に正解と不正解の比が最大となった。CpNpG アイランドに関して、シロイヌナズナでは GC 含量が 50%、長さが 200bp、観測値/期待値が 1.5、イネでは GC 含量が 70%、長さが 300bp、観測値/期待値が 0.6 の時に正解と不正解の比が最大となった。

正解と不正解の比がそれぞれ最大になった指標の閾値を利用して、再度 CpG/CpNpG アイランドを検出すると、CpG アイランドに関して、TSS 周辺[-100,200]におけるシロイヌナズナの正解セットで 498 個、不正解セットで 104 個、イネの正解セットで 122 個、不正解セットで 104 個確認できた。CpNpG アイランドに関して、TSS 周辺[-100,200]におけるシロイヌナズナの正解セットで 43 個、不正解セットで 17 個、イネの正解セットで 122 個、不正解セットで 100 個確認できた。以上のことから、イネにおいても TSS 周辺で CpG と CpNpG の頻度が高くなっていることが確認できた。次にこの CpG/CpNpG アイランドがどの領域に存在するのか観察した（図 2.11、図 2.12）シロイヌナズナでは TSS の上流 100bp から TSS 下流数百 bp にかけて、CpG アイランドが検出されている。既存の研究で報告されているように、シロイヌナズナのプロモーター領域で CpG アイランドは検出されるが、プロモーター領域よりむしろ UTR 領域で検出されることが新たに明らかになった。一方、イネでは TSS の下流 300bp 以降の領域で CpG アイランドが検出されている。プロモーター領域には全く検出されなかった。

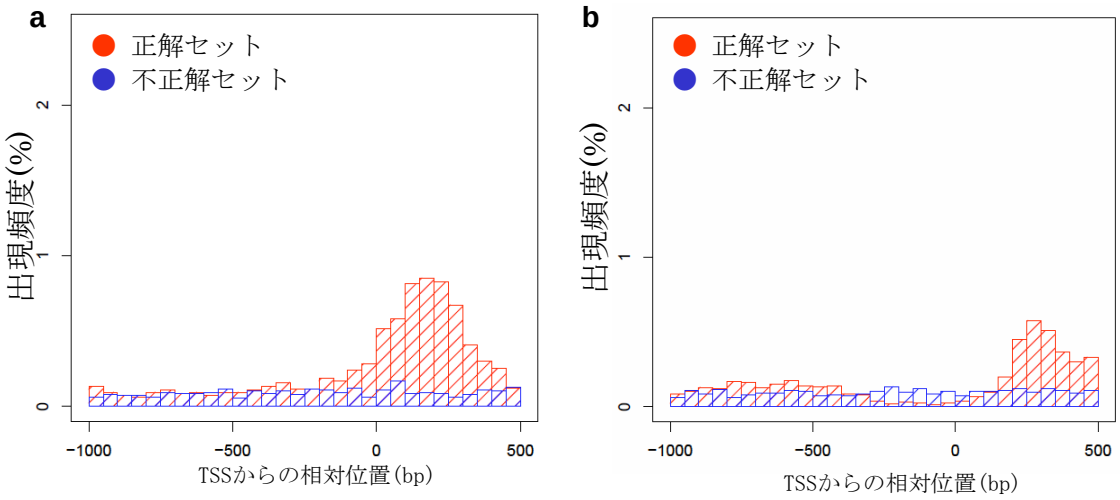

図 2.11 CpG アイランドの出現位置。横軸は CpG アイランドの出現頻度(%)、縦軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

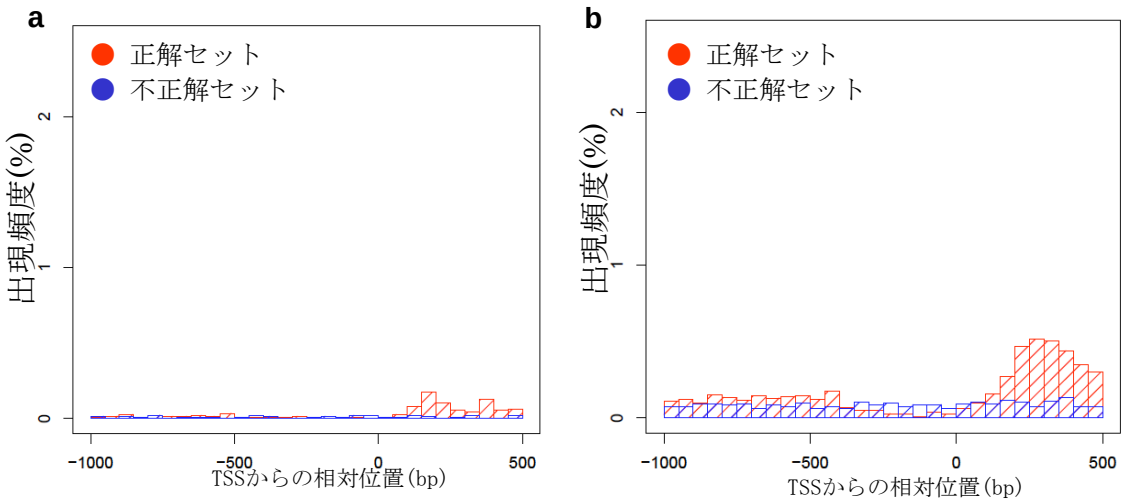

図 2.12 CpNpG アイランドの出現位置。横軸は CpG アイランドの出現頻度(%)、縦軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

シロイヌナズナでは CpG アイランドと同様に CpNpG アイランドでもプロモーター領域を示唆できるという報告がされていたが、今回の結果をみると、プロモーター領域よりUTRやその下流の遺伝子領域のN末端周辺に観察された。イネに関しても同様に、CpG/CpNpG アイランドのどちらもプロモーター領域では検出されず、シロイヌナズナと同様に UTR やその下流の遺伝子領域の N 末端周辺では観察された。

以上のことから、CpG アイランドと CpNpG アイランドは種間で共通して UTR やその下流の遺伝子領域のN末端周辺に存在すると言える。このことから、CpG/CpNpG アイランド何らかの形で植物の転写制御に影響を与えている可能性は十分にある。

2.3.5 情報エントロピー

これまでは主にシロイヌナズナとイネの間でプロモーター領域の特徴を比較してきた。ここでは種内での多様性又は共通性を確認するために、プロモーター領域周辺におけるエントロピーの推移を観察した。繰り返し配列や k-mer の頻度解析によって 2 塩基や 3 塩基の単純な配列が TSS 周辺で位置特異的に存在することが観察されている。ここでは tri-mer の出現頻度によって、正解/不正解セット中の各領域のエントロピーを求めた（図 2.13）。これにより、プロモーター領域における共通性の高い領域を抽出することが出来る。

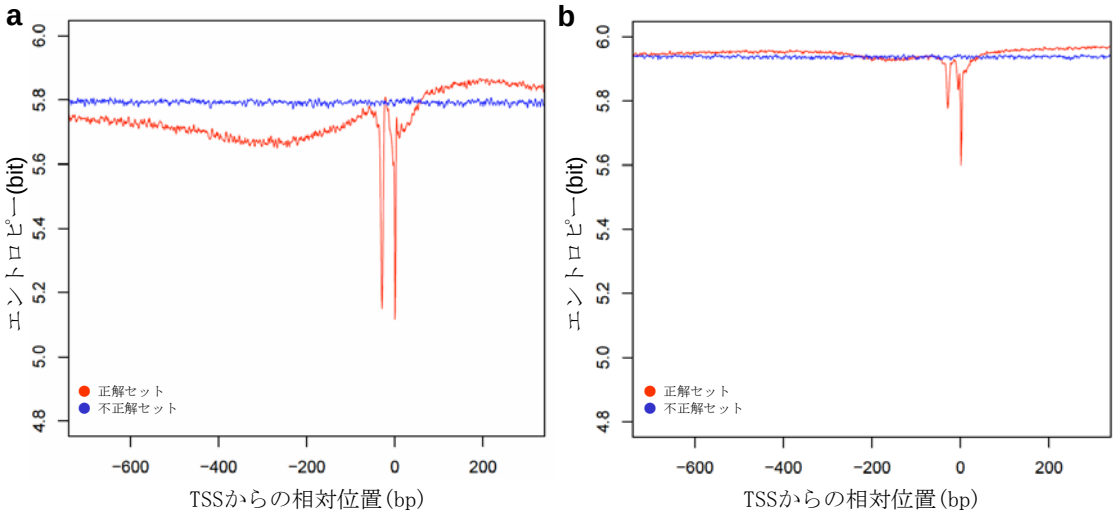

図 2.13　エントロピーの推移。横軸はエントロピー(bit)、縦軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

２種とも TSS 上流 30bp と TSS にそれぞれ TATAbox と Initiator と考えられるエントロピーの減少のピークを確認できた。シロイヌナズナにおいては、TSS 上流の数百 bp に渡ってエントロピーが低い領域を観察した。このことからプロモーター領域の共通性が高いことが分かる。この領域には共通な発現制御配列が存在する可能性がある。イネにおいては、不正解セットと同程度に低い場所が TSS 上流に存在することが観察された。なお、ゲノム配列は完全にランダムな配列が存在しているわけではなく、あるパターンを持つ配列が並んでいる。このことから、各領域のエントロピーが 6 以下を推移している。イネでは TSS 上流に異なるパターンを持つ配列が混在する可能性が示唆された。実際に、後の解析において、イネのデータセットが異なる 4 つのグループに分けられることが確認された (3.3.4)。シロイヌナズナとイネの２種ともに TSS の下流でエントロピーが大きい理由は遺伝子コード領域（CDS）による可能性が考えられる。実際に、UTR の長さを確認したところ、シロイヌナズナでは平均 137.2bp、中間値 98bp、イネでは平均 267.4bp、中間値 125bp という値になった。このことから、シロイヌナズナに関しては CDS の領域によってエントロピーが増加している可能性が高い。イネに関しては UTR 領域でもエントロピーが増加しているようだ。後の解析にて 4 つのグループに分けた時に、各クラスでの UTR 領域のエントロピーが低くなっていたため、実際にイネの UTR には他の領域に比べて異なるパターンの配列が存在していると考えられる。

2.4 まとめ

本章では、植物のプロモーター領域周辺もしくは TSS 周辺に存在すると報告のあった繰り返し配列や CpG/CpNpG アイランドの解析を通して、種間での共通性を確認した。例えば、C と T のみから成る２～３塩基の繰り返し配列が TSS 周辺で多いこと、CpG アイランドと CpNpG アイランドは UTR やその下流の遺伝子領域の C 末端周辺に存在することである。これらは植物のもつ転写因子が共通して認識をする特徴であると考えられる。

シロイヌナズナでは繰り返し配列の有無で正解セットを分けても GC-skew や GC 含量の推移がほとんど変わらず、エントロピーの値も低かった。このことからプロモーター領域の共通性が高いことが考えられる。イネでは繰り返し配列の有無で GC-skew や GC 含量の推移が異なること、エントロピーの値が高いことからシロイヌナズナのプロモーター領域に比べて多様な特徴をもつことを明らかにした。

動物に比べ多くの転写因子を持っている植物においてはプロモーター領域も多様であると考えていたにもかかわらず、シロイヌナズナのプロモーター領域の共通性が高いと言うことは興味深い結果と言える。このプロモーター領域の共通性と転写因子の増加を考えると、一つのプロモーター領域と多数の種類の転写因子が結合すると考えられる。一方で、動物（哺乳類）は植物に比べると転写因子が少なく、さらにオルタナティブプロモーターが存在することも確認されている（Suzuki *et al* 2002、Landry *et al* 2003、Kimura *et al* 2006）。このことは、動物は一つのプロモーター領域と少ない種類の転写因子が結合していると考えられ、プロモーター領域に結合する転写因子の特異性が高いと考えられる。

今後、完全長の cDNA の解析が進み、データが蓄積されれば、単子葉植物及び双子葉植物としての共通性や多様性を明らかすることや、植物と動物とを比較することで高等生物としての共通性や多様性を明らかにすることも可能になるだろう。

第３章

プロモーター領域の予測

3.1 背景と目的

近年の完全長の cDNA の解析によって、正確かつ大量にプロモーター領域を同定することが可能になり、それに伴い特徴解析も行われている。次なる目的プロモーター領域の予測を行い、その予測精度を検証することで新たな知見を得る試みを行う。

プロモーター領域の予測は哺乳類や酵母で研究が進んでいる。特に先に述べた CpG アイランドによって哺乳類のプロモーター領域の予測プログラムの精度は大きく改善した(Davuluri *et al* 2002、Down and Hubbard 2002、Bajic *et al* 2003、Bajic *et al* 2004)。一方で植物に特化したプロモーター領域の予測プログラムは Shahmuradov らによって 2005 年に発表された（Shahmuradov *et al* 2005)。この研究では、PlantProm と呼ばれる文献から得られたプロモーター領域のデータベース（単子葉植物から 71 個、双子葉植物から 220 個、その他の植物から 14 個、計 305 個）を利用している。このプログラムでは TATAbox の有無でモデルを分け、CpG アイランドや転写因子結合領域のモチーフといった特徴を利用し、Support Vector Machine を使って予測を行っている。このプログラムではシロイヌナズナの 13,350 個の遺伝子のタンパク質コード領域(CDS)を利用して、その上流 5000bp の領域までを探索し、9,633 個を正解と予測した（感度 72.3%）と報告されている。

このプログラムの問題点として２つの点が上げられる。１つ目は多種多様な植物のプロモーター領域の予測を目的としているにもかかわらず、シロイヌナズナでしか有効性の報告が無い CpG アイランドを予測に組み込んでいる点である。２つ目はすべての植物のプロモーター領域の予測を共通のモデルで予測しようとしている点である。

この章では、特徴解析によって得られた知見を利用して、プロモーター領域の予測を行う。また既存の研究では検討がされなかった植物のプロモーター領域を一つのモデル

で予測できるのかという問題について、一つの生物によって作ったモデルを他の植物のデータに適応することで検証を行う。また、これらの解析を通して、種間及び種内でのプロモーター領域の特徴の共通性や多様性を明らかにしていく。

3.2 方法

3.2.1 データセット及び解析対象領域の抽出法

シロイヌナズナとイネのデータセットは特徴解析に作成したものを使用した（2.2.1)。これまでと同様の定義により、プロモーター領域のデータセットを正解セット、プロモーター領域以外のデータセットを不正解セットとした。ここでは、上記の２種に加えてタルウマゴヤシのデータセットも作成した。cDNA 配列を MtGI (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=medicago)から、ゲノム配列を IMGAG（http://www.medicago.org/genome/IMGAG/）から、それぞれダウンロードして利用した。MtGI は EST 配列から作成した cDNA 配列が格納されている。タルウマゴヤシは完全長のゲノム配列の解読が完了しておらず、IMGAG には BAC の配列が格納されている。特徴解析の場合と同様に cDNA 配列をゲノム配列にマップして TSS を決定する。決定した TSS から上流 1000bp から 下流 500bp の領域[-1000,+500]を正解セット（プロモーター領域）とし、[-2000,+1000]以外の領域から長さ 1501bp の領域を抽出し不正解セットとした。それぞれ 1793 個の正解セットと不正解セットを作成した。

3.2.2 評価方法

予測の評価は 10 分割交差検定により行った。予測精度の評価指標としては感度(Sn)、特異度(Sp)、相関係数(Cc)を用いた。それぞれの定義は以下の通りである。

$$\text{感度}(Sn) = \frac{TP}{TP + FN} \tag{3.1}$$

$$\text{特異度}(Sp) = \frac{TN}{TN + FP} \tag{3.2}$$

$$\text{相関係数}(Cc) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.3}$$

ここで、*TP* は正解と予測したもののうち実際に正解のもの、*TN* は不正解と予想したもののうち実際に不正解のもの、*FP* は正解と予測したもののうち実際は不正解のもの、*FN* は不正解と予測したものうち実際は正解のものである。

また、他の予測精度の指標として受信者動作特性（ROC）曲線を用いる。ROC 曲線とは縦軸に感度、横軸に（１－特異度）をプロットし、閾値（カットオフポイント）を媒介変数として変化させた時に描ける曲線のことである。単位正方形の中で ROC 曲線よりも下側の領域の面積、Area Under the Curve（AUC）は、正答率とみなせることが証明されている（Green and Swets 1966)。求められた ROC 曲線から AUC を求めて予測精度の一つの指標とする。

3.2.3 ２次のマルコフモデルを使った予測

１次のマルコフモデルを利用した判別（阿久津ら 2001）を２次のマルコフモデルに改良し予測を行った。具体的なアルゴリズムは以下の通りである。

有限個の状態の集合 $S=\{s_1,s_2,\cdots,s_m\}$を値にとる確率過程の状態時系列 $x=x_1x_2\cdots x_T$ において、任意の時点 t での状態 x_t がそれ以前の２個の状態系列 $x_{t-1}x_{t-2}$ だけを条件として決定すると仮定すると、ある状態が別の状態に移行する確率（遷移確率）は次の式にて表すことができる。

$$a_{s,u} = P(x_t = u \mid x_{t-1}x_{t-2} = s) \qquad (3.4)$$

ゲノム配列に関する任意の確率モデルについて $P(X,Y)=P(X|Y)P(Y)$を何度も適応することによって、配列の確率は次のように表すことが出来る。

$$P(x) = P(x_L, x_{L-1},..., x_1)$$
$$P(x_L \mid x_{L-1},..., x_1)P(x_{L-1} \mid x_{L-2},..., x_1),..., P(x_1)$$

２次のマルコフ過程を仮定していることから、各シンボル x_i の確率が先行するシンボル $x_{i-1}x_{i-2}$ にだけ依存し、それ以前に現れた配列には依存しない。それゆえ、式(3.4)を使って前出の式は以下のようになる。

$$P(X) = P(x_L \mid x_{L-1}x_{L-2})P(x_{L-1} \mid x_{L-2}x_{L-3}),..., P(x_2 \mid x_1)P(x_1)$$

$$P(x_1)P(x_2 \mid x_1)\prod_{i=3}^{L} a_{x_{i-1}x_{i-2},x_i} \tag{3.5}$$

式(3.5)を使って最尤比検定を行う。まず、正解セット(+)と不正解セット(-)各々の遷移確率を次式によって求める。

$$a^{+}_{s,u} = \frac{C_{s,u}}{\sum_{u'} C^{+}_{s,u'}} \tag{3.6}$$

a^{-}についても同様に求める。ここで$C_{s,u}$は任意の地点 t における配列 s に続く配列 u の出現回数である。これらの正解セット(+)と不正解セット(-)のモデルを判別に利用するために対数オッズ比を計算する。

$$S(x) = \log \frac{p(x \mid model+)}{p(x \mid model-)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_{i-2},x_i}}{a^{-}_{x_{i-1}x_{i-2},x_i}}$$

$$= \sum_{i=1}^{L} \beta_{x_{i-1}x_{i-2},x_i} \tag{3.7}$$

ここで x は配列、$\beta x_{i-1}x_{i-2},x_i$は対応する遷移確率の対数尤度比である。このスコア $S(x)$ の値によって判別を行う。

3.2.4 重回帰分析

先の遷移確率の対数尤度比$\beta x_{i-1}x_{i-2},x_i$に対して、各領域 i の予測へ影響度の違いにより重み付けし、予測精度の向上を図る。ここでは、重みを重回帰により決定する。目的変数(Y)に正解セットでは１、不正解セットでは－１を与え、$\beta x_{i-1}x_{i-2},x_i$を説明変数とした。回帰式は以下のようになり、この係数 k_i を最小二乗法で決定する。なお、各領域のスコアは独立なものであると仮定している。

$$Y = \sum_{i=1}^{L} k_i \beta_{x_{i-1}x_{i-2},x_i} \tag{3.8}$$

この係数 k_i は各領域の重要度を示すもので、この値が高い領域ほど予測時に与える影響が強い。

3. 2. 5 クラスタリング

クラスタリングとはデータセットを互いに類似したデータごとにグループ分けすることである。2.3.5 にてイネの正解セットはシロイヌナズナの正解セットに比べて多様性が存在することが確認されている。多様性の高いデータセットを一つのモデルで予測するより、データを各 DNA 配列の類似度にて分類することで予測精度の向上が期待できる。具体的には、配列中から k-mer の出現頻度によって 4 の k 乗次元の変数を抽出してデータセットを作成する。これらのデータセットを使用し、非階層的クラスタリングである K-means クラスタリングを行った。

3. 2. 6 Gene Ontology (GO)

イネのプロモーター領域のもつ特徴と遺伝子の機能について検討するためにイネの遺伝子にアノテーションされている GO を利用した。今回の解析で使用した TIGR のデータベースには GO のアノテーションがついていないため、KOME(http://cdna01.dna.affrc.go.jp/cDNA/)のデータベースを利用し、TIGR のデータに対して GO のアノテーションをつけた。KOME には TIGR と同様に完全長 cDNA が格納されており、GO によるアノテーションがついている。この KOME の配列に TIGR の配列を BLAST にかけて、最も相同性が高いものを選択し、TIGR の配列に GO のアノテーションをつけた。なお、相同性が 95%以下のものは取り除いた。

分類する term の選択は Kikuchi らの行ったイネの完全長 cDNA の論文を参考に決定した(Kikuchi *et al.* 2003)。具体的には、1.代謝(Metabolism)、2.輸送(Transport)、3.転写(Transcription)、4.翻訳(Translation)、5.エネルギー(Energy)、6.発生(Development)、7.局在(Localization)、8.刺激(Stimulus)、9.細胞死(death)、の 9 つ GOterm を採用した。

以上、特にプログラムの明記が無い場合には、perl、java、R、を使用して解析を行った。

3.3 結果と考察

3.3.1 マルコフモデルを使った予測

繰り返し配列や k-mer の出現頻度の解析によって、di-mer、tri-mer といった単純な配列の頻度がプロモーター領域周辺で位置特異的に変化していることが分かった（第２章)。そこで、低次のマルコフモデルを使うことによってこれらの特徴を捕らえることが出来ると考え、２次のマルコフモデルを使ってプロモーター領域の予測を行った。予測時の学習領域が予測精度に違いを生むため、まず学習領域を決定する必要がある。そこで、TSS から上流 500bp から下流 200bp まで領域[-500,200]を検証する学習領域と定め、その領域内で上限と下限を 100bp ずつずらし、ウィンドウ幅も変えて予測精度の検証を行うことした。この手法をシロイヌナズナとイネのデータセットにそれぞれ適応し、学習領域と予測精度について検証を行った（図 3.1、3.2)。

シロイヌナズナでは[－300,200]の領域を、イネでは[－400,100]を学習領域として予測に組み込むことで予測精度（AUC）が最大となった。UTR の領域を組み込むことで予測精度が改善されていることが分かる。予測精度が最大値を示したときの学習領域を利用して予測した結果を ROC 曲線にて示す（図 3.3)。相関係数(*Cc*)が最大になるように閾値を設定するとシロイヌナズナで *Sn*＝88.66%、*Sp*= 89.63%、*Cc*= 70.73%、イネで *Sn*=83.83%、*Sp*= 64.31%、*Cc*＝44.90%という結果となった。

イネの予測精度はシロイヌナズナと比べて低い。この結果は、イネではプロモーター領域にシロイヌナズナと比べて多様性がみられたこと（2.3.5）に起因するものと考えられる。

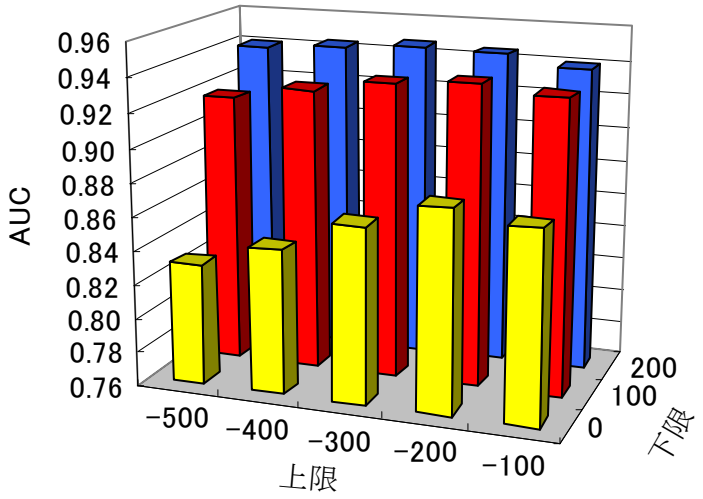

図 3.1 学習領域と予測精度の関係（シロイヌナズナ）

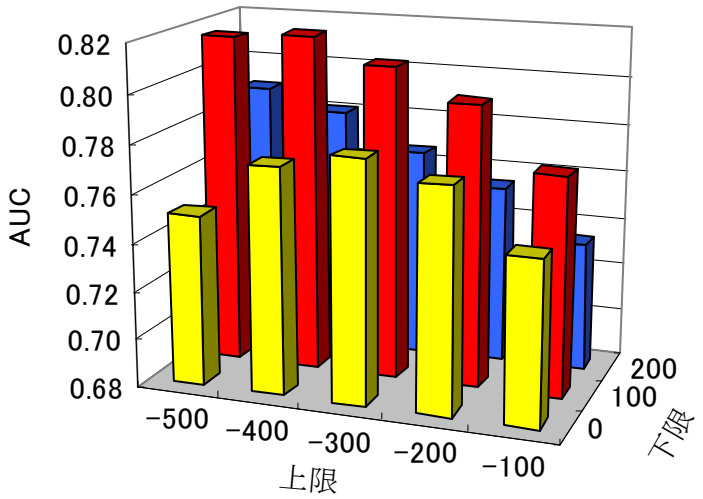

図 3.2 学習領域と予測精度の関係（イネ）

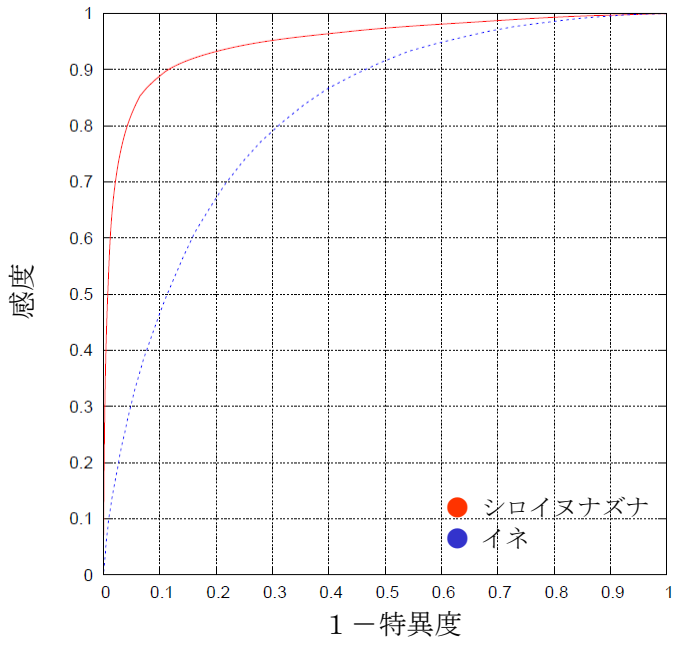

図 3.3 ２次のマルコフモデルを使った予測。縦軸は感度、横軸は 1-特異度。

3. 3. 2 重回帰分析を併用した予測

ここでは、２次のマルコフモデルを使った予測から、予測精度を高めることを検討する。これまでの k-mer やエントロピーの解析によって、プロモーター領域の周辺では共通性の高い領域と低い領域が存在することが確認された（2.3)。例えば、TSS 周辺には Initiator、TSS 上流 30bp には TATAbox という特徴的な配列が存在する(2.3.5)。これらの領域は正解セットの中で比較的共通性が高く、他の領域に比べて強くプロモーター領域を示唆することが考えられる。そこで各領域に与えられたスコア（遷移確率の対数尤度）に対して重回帰分析により重みをつけ、予測を行った。シロイヌナズナとイネのデータセットにそれぞれ適応し、既存の２次のマルコフモデルとの予測精度の比較を行った（図 3.4)。

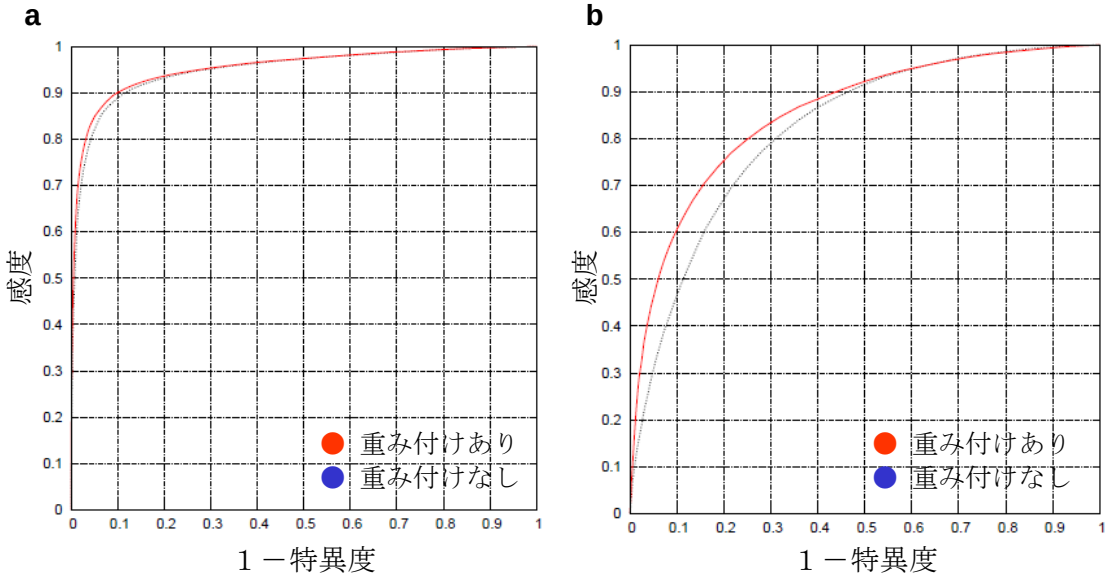

図 3.4 重回帰分析を使った重み付けの有無による予測精度の違い。縦軸は感度、横軸は 1-特異度。**a** シロイヌナズナ。 **b** イネ。

２種ともに予測精度の改善が確認された。相関係数(*Cc*)が最大になるように閾値を設定したときの予測精度を示す（表 3.1、表 3.2)。シロイヌナズナでは感度を 0.72 ポイント、特異度を 1.41 ポイント上昇させることに成功した。イネでは感度を 0.8 ポイント、特異度を 3.8 ポイント上昇させることに成功した。

表 3.1 重回帰分析を使った重み付けによる予測精度の改善（シロイヌナズナ)

	Sn(%)	*Sp*(%)	*Cc*(%)
重み付けなし	88.66	89.63	70.72
重み付けあり	89.38	91.04	73.22

表 3.2 重回帰分析を使った重み付けによる予測精度の改善（イネ）

	Sn(%)	*Sp*(%)	*Cc*(%)
重み付けなし	83.83	64.31	44.90
重み付けあり	84.63	68.11	48.05

次に、実際にどの領域に対して高く重み付けがされているのか確認するために、重回帰分析によって得られた各領域の重みを示す（図 3.5)。TSS 周辺で重みは増加し、上流 30bp 辺りでも増加していることが２種で確認できた。この結果は（2.3.5）に述べた結果に矛盾しない。また、これらの重みの推移はエントロピーの推移と類似しており、もっともらしい重み付けがされていると考えられる。

以上の結果から、２次のマルコフモデルと重回帰分析による位置の重み付けを組み合わせた提案手法はシロイヌナズナとイネのプロモーター領域の予測に有効であると考えられる。

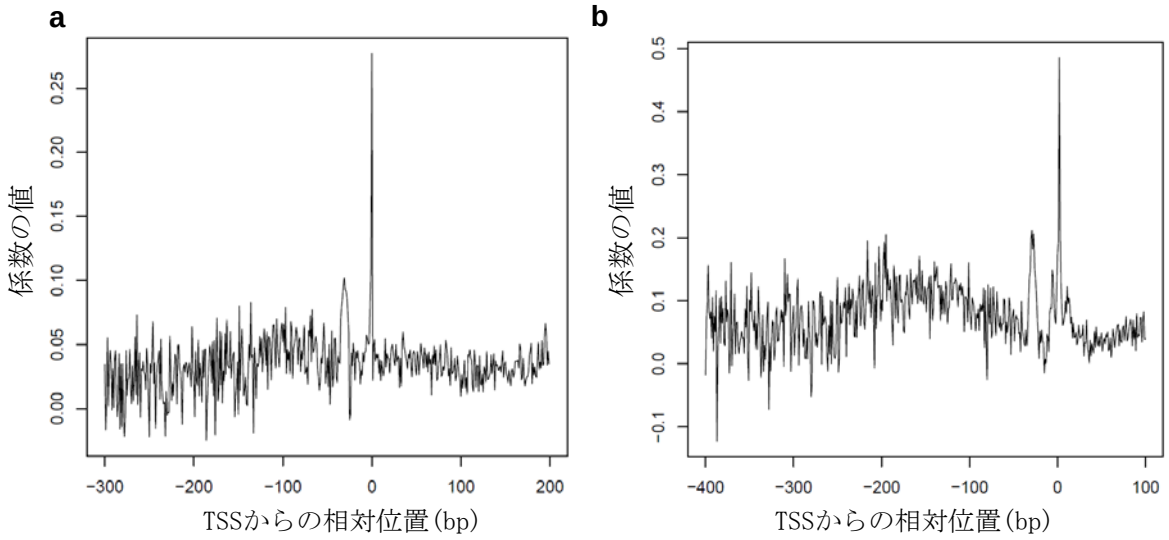

図 3.5 重回帰分析による重みの推移。縦軸は係数の値、横軸は TSS からの相対位置(bp)。**a** シロイヌナズナ。**b** イネ。

3.3.3 クラスタリングを併用した予測

3.3.1、3.3.2 に述べた予測では、イネの予測精度はシロイヌナズナと比べるとかなり低かった。ここでは、主にイネについて重み付けとは異なる方法で予測精度の向上を試みる。イネでは、全体の領域を通してエントロピーの値が高いことや、繰り返し配列の有無で GC-skew や GC 含量に大きな違いがあることが分かっている（2.3）。このことから、シロイヌナズナと比べて、プロモーター領域の塩基組成に共通性が低いことが推測できる。そこで、すべてのデータセットを一つのモデルとして学習するのではなく、いくつかのクラスに分けてそれぞれ別のモデルで学習することを検討する。このことによって各クラス内での塩基組成の共通性が高まり、予測精度が改善すると期待される。2 次のマルコフモデルを使用していることから、ここでは 3-mer の出現頻度（64 次元の変数）に基づき K-means によりクラスタリングを行う。

K-means クラスタリングにおいては、そのクラスタ数についてまず検討する必要がある。ここでは、クラスタリング後のデータセットにおいて 2 次のマルコフモデルを学習させ、10 分割交差検定において予測精度が高いものほど良い分類であるとし、そのときの K の値を採用することにした。また、クラスタリングの際に必要な変数（3-mer の出現頻度）を作成する時に利用する領域も決定しなければならない。その領域によって、クラスタリングの精度は大きく異なってくる。例えば、バックグラウンドと変わらない特徴を持つ領域と、TSS や UTR のように繰り返し配列の有無で GC 含量や GC-skew の推移の違いがはっきり観察できた領域では後者の領域を使う方がデータセットをうまく分類できるはずである。

以上のことから、まずはイネのデータセットにおいてクラスタ数とクラスタリング時に使用する学習領域と予測精度について検証を行った（図 3.4）。クラスタ数は 2 ～ 5 個、使用する領域は TSS の上流 500bp から下流 200bp まで上限と下限を 100bp ずつずらして予測精度の検証を行った。同様にシロイヌナズナのデータセットでも検証を行った（図 3.5）。

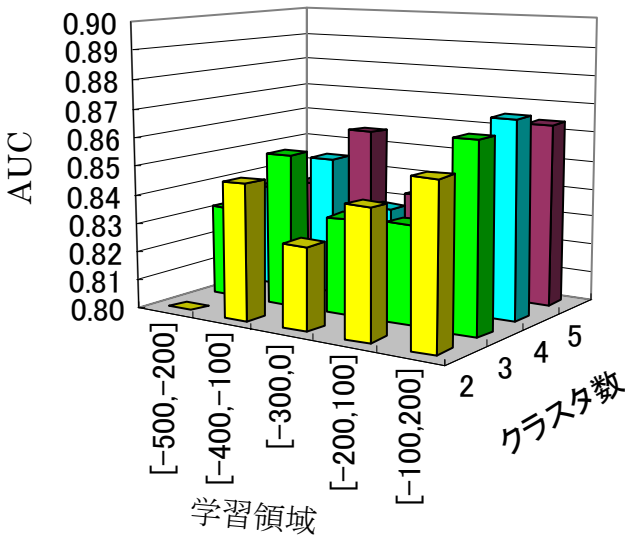

図 3.4 クラスタ数と使用する領域の違いによる予測精度の違い（イネ）

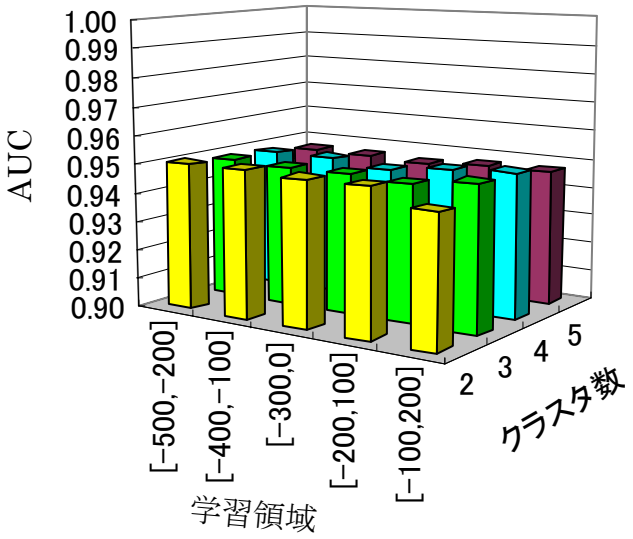

図 3.5 クラスタ数と使用する領域の違いによる予測精度の違い（シロイヌナズナ）

イネにおいては、クラスタ数が 4 でクラスタリング時に使用する領域が[-100,200]のときに AUC の値が 0.87 となり、予測精度が最大になった。この領域は繰り返し配列が多数存在した領域や、エントロピーの減少が見られた領域と一致している。イネのすべてのデータセットを使った時の予測では AUC の値が 0.82 であったことから、クラスタリングを併用することで予測精度が改善された。特異度にして約 10 ポイントの上昇が確認できた。

シロイヌナズナにおいては、クラスタ数が 2 でクラスタリング時に使用する領域が[-400,-100]のときに AUC の値が 0.95 となり、予測精度が最大になった。シロイヌナズナのすべてのデータセットを使った予測では AUC の値が 0.95 であったことから、クラスタリングを併用しても予測精度はほとんど改善されなかった。特異度にして約 0.6 ポイントの上昇に過ぎなかった。このことは、シロイヌナズナの場合はプロモーター領域の共通性が高いという結果（2.3.5）に矛盾しない。よって、シロイヌナズナのデータセットに関しては一つのモデルで予測したほうが良いと考えられる。

3.3.4 重回帰分析とクラスタリングを組み合わせた予測

イネに関してはクラスタリングを使った予測と重回帰分析を使った予測のどちらも予測精度の向上に成功した。これらの予測を組み合わせることで、イネの予測精度がどの程度まで高くなるか検証した（図 3.6）。具体的には、クラスタリング後の各データセットにおいて重回帰分析による重み付けを行った。ROC 曲線をみると、予測精度が向上したことが分かる。クラスタリングを使った予測、重回帰分析を使った予測、それらを組み合わせた予測について、閾値を相関係数（Cc）が最大になるように設定した時の予測精度を整理したものを表 3.3 に示す。2 次のマルコフモデルのみ使った予測と比較して、重回帰分析とクラスタリングを併用した場合は感度が 0.8 ポイント、特異度が 12.24 ポイント上昇した。

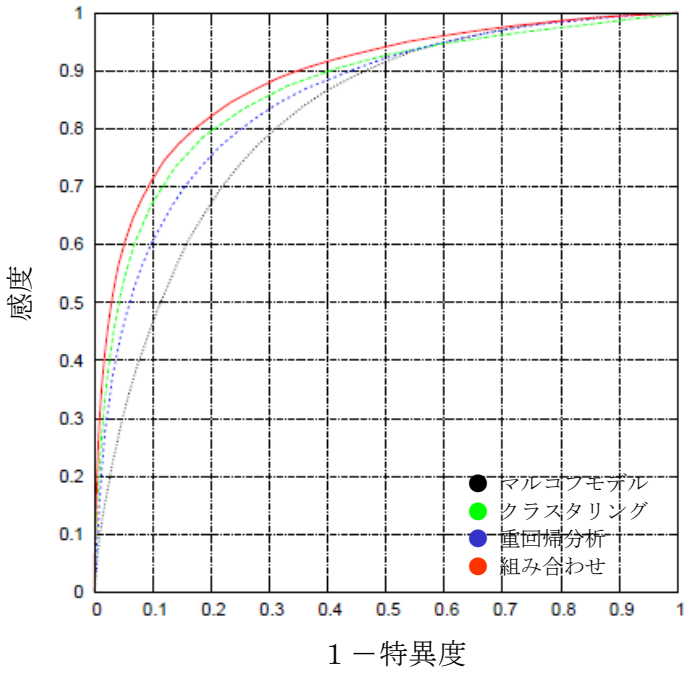

図 3.6 各種予測の予測精度の改善。縦軸は感度、横軸は 1-特異度。

表 3.3 改良された予測精度

	Sn(%)	*Sp*(%)	*Cc*(%)
マルコフモデル	83.83	64.31	44.90
重回帰分析を併用	84.63	68.11	48.05
クラスタリングを併用	83.52	74.34	51.02
組み合わせ	84.63	76.55	53.90

相関係数(*Cc*)が最大になるように閾値を設定している。

3.3.5 クラスタリングによって得られた各クラスの特徴

先の解析で、イネのデータセットを TSS 周辺の 3-mer の出現頻度によって４つのクラスに分類することで、予測精度を向上できることを明らかにした。このことはイネのプロモーター領域がシロイヌナズナにない多様性を持っていることを示唆している。ここでは、これらの各クラスにどのような特徴が存在するかを明らかにするために、イネについて、各クラスの GC 含量、GC-skew、エントロピーについて解析を行った。また、特徴の違うプロモーター領域はそれぞれどのような機能を持つ遺伝子の発現に関わるのかを明らかにするために、遺伝子にアノテーションされた GO を利用し、各クラスの違いを観察した。

3.3.5.1 各クラスの GC 含量と GC-skew

まず、各クラスで TSS からの相対位置による GC 含量や GC-skew の推移を観察した（図 3.7、3.8）。ここで各クラスに対して１から４までの番号をつけた。クラス１の個数は 8205(43.81%)、クラス２の個数は 3297(17.60%)、クラス３の個数は 3312(17.68%)、クラス４の個数は 3917(20.91%)である。

クラス１では TSS 周辺の GC-skew の増加が観察できず、GC 含量の推移も他のクラスのプロモーター領域の特徴とは大きく異なっている。さらにその数は全体の 43.81%と多い。このような推移をもつプロモーター領域がシロイヌナズナでも観察できるか確かめるために、先ほどのクラスタリング手法でシロイヌナズナのデータセットを２～５個のクラスに分けて、それぞれで GC 含量と GC-skew の推移を観察したが、イネのクラス１のようなものは観察できなかった。このことから、イネのクラス１はシロイヌナズナにはほとんど存在しないプロモーター領域であると考えられる。その他のクラスでは、クラス３で TSS 周辺の GC-skew が強く増加していることや、クラス２と３で UTR 領域の GC 含量が強く増加するなど、各クラスでそれぞれ異なる特徴が存在することを確認した。

GC 含量に注目すると、クラスタリングに使用した領域[-100,200]から離れた領域[-600,-200]においても GC 含量の推移に違いを観察した。このことは TSS 周辺の[-100,200]の領域と TSS 上流[-600,-200]の領域には何らかの関係があることを示唆している可能性がある。

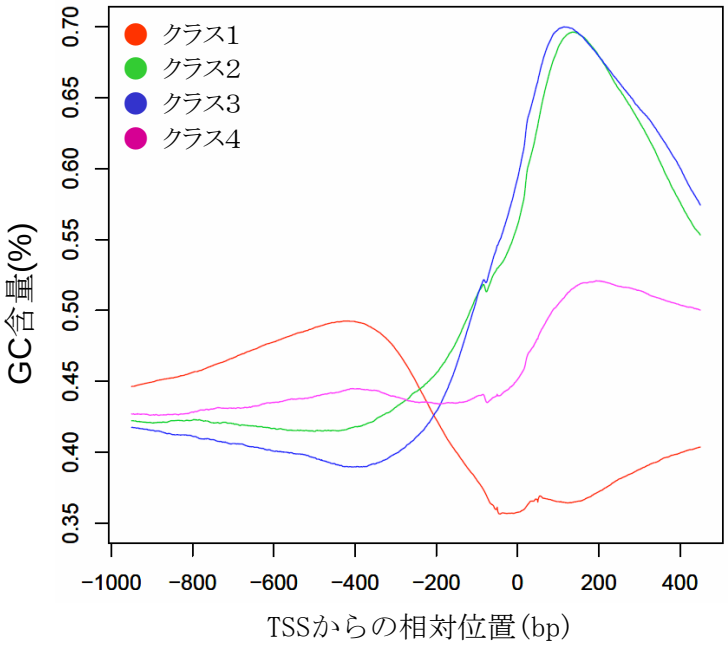

図 3.7 各クラスでの GC 含量の推移。縦軸は GC 含量(%)、横軸は TSS からの相対位置(bp)。

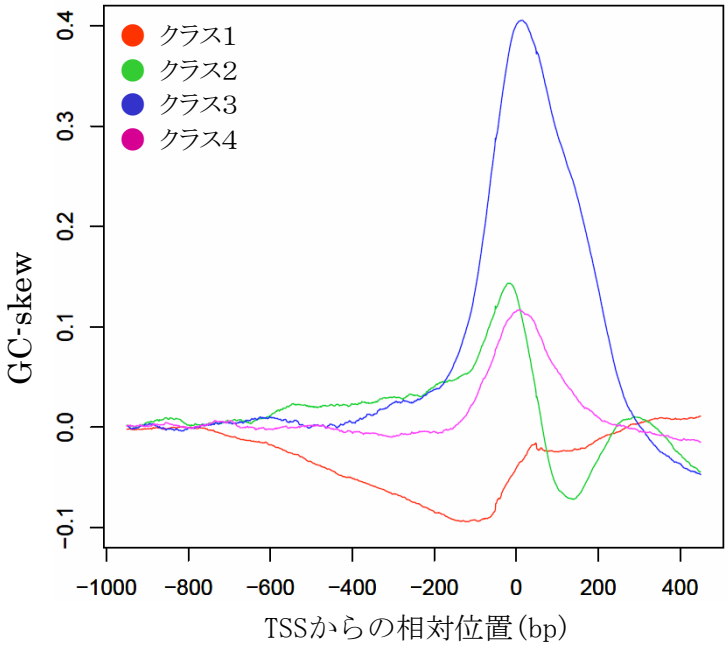

図 3.8 各クラスでの GC-skew の推移。縦軸は GC-skew、横軸は TSS からの相対位置(bp)。

3.3.5.2 各クラスの情報エントロピー

GC 含量や GC-skew といった統計量には各クラス間で違いが分かること分かった。次に、エントロピーを確認することで、各クラス内の共通性を確認する。また、すべてのデータセットを使った時のエントロピーと比較し、各クラス内での共通性が高くなったかどうか確認する。正解/不正解セット中の各領域のエントロピーを求め、各クラス内での塩基組成の多様性と共通性を観察した（図 3.9）。

どのクラスでも TSS から UTR にかけて、正解セットのエントロピーの減少が確認できる。各クラス内の正解セットと不正解セットの差を見ても、クラス３に例外が見られるものの、イネのすべてのデータセットを使った場合（図 2.13）より大きくなっていることが分かる。このことからクラスタリングを行うことによってクラス内の共通性が高まり、予測精度の向上に繋がったといえる。

クラス１に関しては他のクラスと違った特徴が確認された。例えば、他のクラスと比べて TSS でのエントロピーの減少が大きいこと、TSS 上流 10 bp 辺りにエントロピーの減少が観察された。また、TSS 周辺[-10,10]のモチーフを各クラスで抽出したところ、クラス１では TTTkyT、クラス２では CkCCnC、クラス３では CnCCnC、クラス４では CTTsCw というモチーフを検出した。このことから、クラス１に関して Initiator のモチーフが他のクラスと全く違うことが分かった。

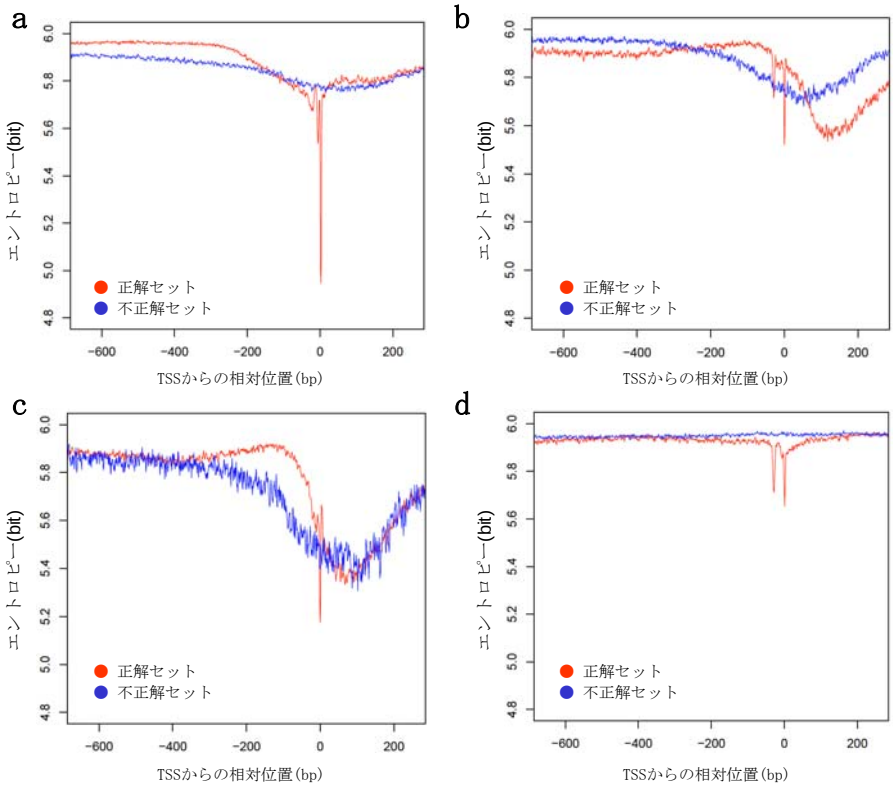

図 3.9 各クラスのエントロピーの TSS からの距離による推移。縦軸はエントロピー(bit)、横軸は TSS からの相対位置(bp)。**a** クラス１。 **b** クラス２。 **c** クラス３。 **d** クラス４。

3. 3. 5. 3 関連遺伝子の GO アノテーション

今回作成した４つのクラスのプロモーター領域のそれぞれに特徴が見られることが分かった。ここでは、各クラスのプロモーター領域とその下流にある遺伝子の機能の関係について解析を試みた。ここでは、GO アノテーションを用いて、各クラスのプロモーター領域と関連する遺伝子の機能を調べた（図 3.9)。GO のアノテーションを少なくとも一つ付けることが出来た遺伝子はクラス１で 3673 個（クラスでアノテーションされた割合は 44.77%)、クラス２で 1494 個(45.31%)、クラス３で 1489 個(44.96%)、クラス４で 1790 個(45.70%)、すべてのクラスで 8469 個(45.21%)であった。

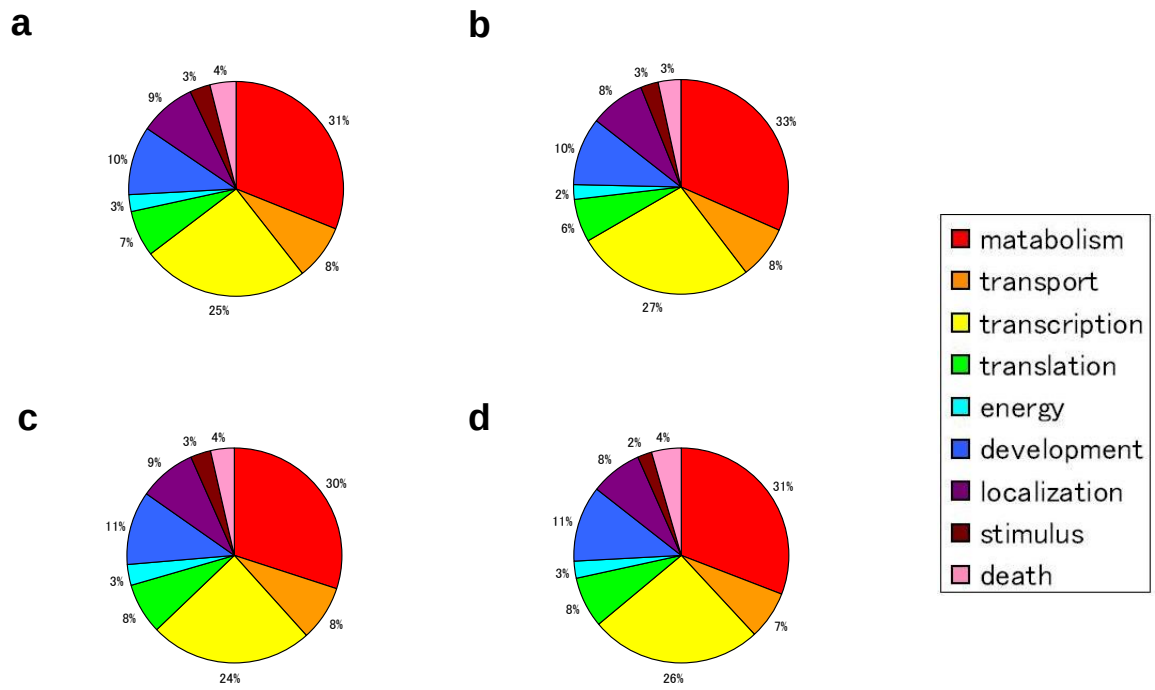

図 3.9 各クラスのプロモーターをもつ遺伝子の GO。**a** クラス１。**b** クラス２。**c** クラス３。**d** クラス４。

結果、各クラスと遺伝子の機能には関連が見られなかった。しかしながら、今回示した GOterm は大きな分類の機能しか選んでいない。特定の機能を表す GOterm を選ぶことで各クラスに違いが確認できる可能性はあるだろう。例えば、特定の機能を表す GOterm を選んだところ、「light-harvesting complex」とアノテーションされた遺伝子の 83 個のうち 53 個（63.86%）をクラス１で確認した。クラス２、３、４ではそれぞれ 8 個、13 個、9 個を確認した。これらは感光性に関与すると示された遺伝子である。シロイヌナズナとイネは長日植物と短日植物であり、開花時期の違う植物である。クラス１のプロモーター領域がシロイヌナズナにはほとんど存在しない GC 含量と GC-skew の推移を示すことから、これらクラス１のプロモーター領域と感光性に関与する遺伝子に何らかの関係がある可能性が考えられる。

3.3.6 他の植物への適用

シロイヌナズナとイネという 2 つの植物のプロモーター領域を同じモデルで予測可能かどうかを検討するため、ここでは高い予測精度を示したシロイヌナズナのモデルを使ってイネのデータセットを予測した。イネのすべてのデータセットと 2.2.3 で作成した４つのクラスの計 5 個のテスト用のデータセットを用意し予測を行った（図 2.13)。クラス２について最も良い予測精度を示した。の精度は比較的良い結果を示した。相関係数（Cc）を最大するように閾値を設定した時のクラス２の予測精度は Sn=90.26%、Sp=52.09%、Cc=53.87%となった。この予測精度はイネのすべてのデータセットで２次のマルコフモデルを使って予測を行った時より高い予測結果となった。一方でクラス１は全く予測できなかった。このことから、これらのシロイヌナズナのプロモーター領域ではほとんど存在しないと考えられるプロモーター領域のデータセット（クラス１）によって、すべてのイネデータセットを使ったときの予測精度が低いと考えられる。

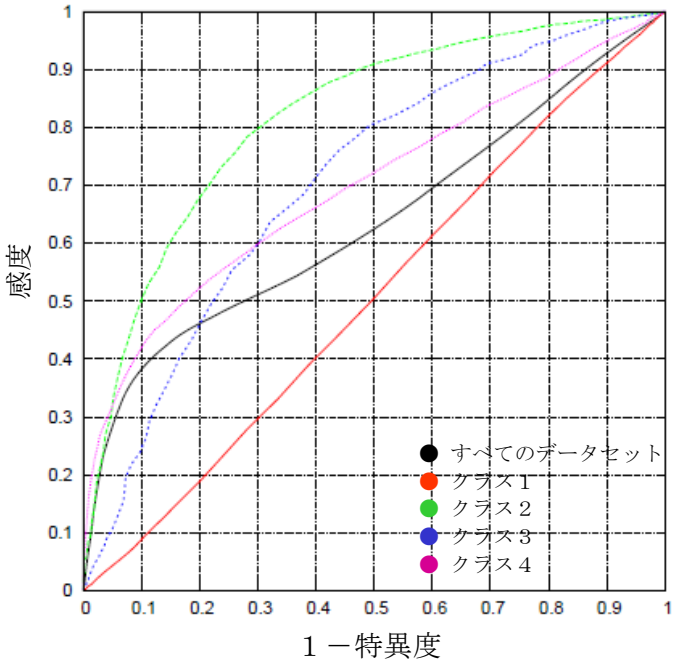

図 3.13 シロイヌナズナのモデルでイネのデータセットを予測した時の予測精度。縦軸は感度、横軸は 1-特異度。

シロイヌナズナのモデルでイネの一部のデータセットに関しては予測が可能であることが分かった。次に、イネよりシロイヌナズナに進化的に近い植物であるタルウマゴヤシのデータセットを作り、シロイヌナズナのモデルでどの程度予測可能なのかの検証を行った（図 3.14）。タルウマゴヤシのデータセットの数は 1793 個である。なお、作成したタルウマゴヤシのデータセット数が少ないのは、ゲノム配列として完全長のゲノム配列ではなく BAC の配列を利用したため、ゲノム配列にマップできた cDNA 配列が少なかったからである。シロイヌナズナのモデルでイネの予測を行った時と比較すると、予想に反し低い予測精度となった。シロイヌナズナやイネが完全長 cDNA の配列によって作られた cDNA 配列なのに対して、タルウマゴヤシのデータが EST 配列によって作られた cDNA の配列を使用しているため、正しいデータセットが作成できていない可能性がある。このことを検証するため、エントロピーと GC-skew の解析、及び 2 次のマルコフモデルによる交差検定を行った。（図 3.15~3.17）

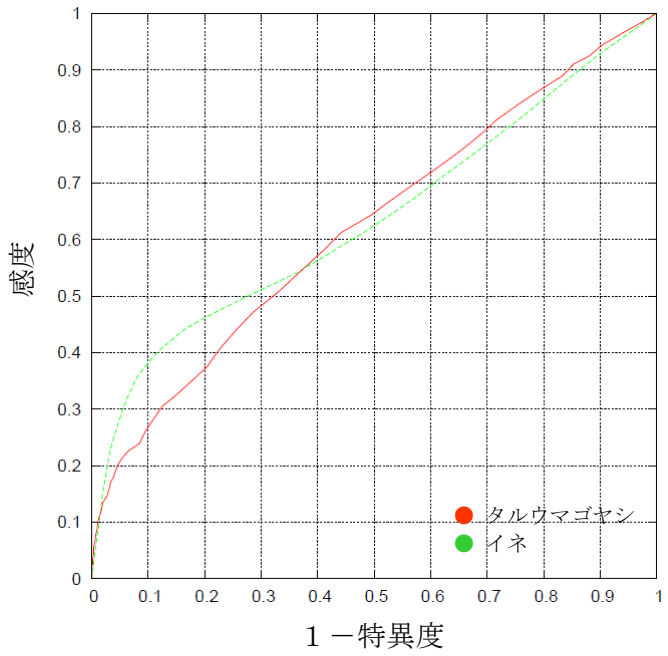

図 3.14 シロイヌナズナのモデルでタルウマゴヤシとイネのデータセットを予測した時の予測精度の違い。縦軸は感度、横軸は 1-特異度。

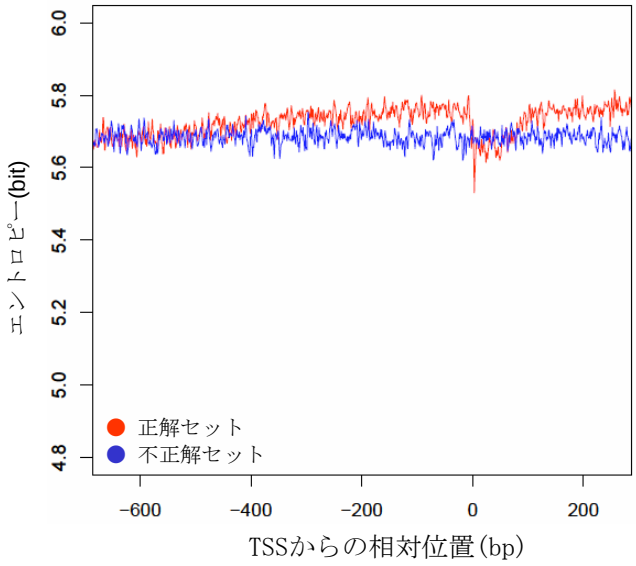

図 3.15 エントロピーの推移。縦軸はエントロピー(bit)、横軸は TSS からの相対位置(bp)。

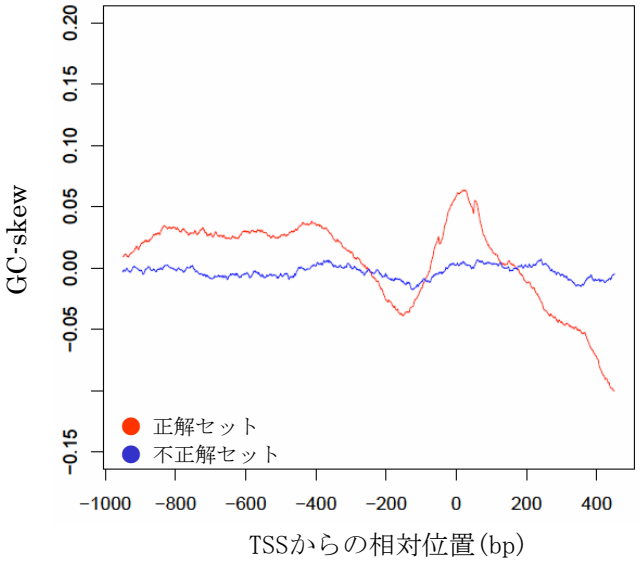

図 3.16 GC-skew の推移。縦軸は GC-skew、横軸は TSS からの相対位置(bp)。

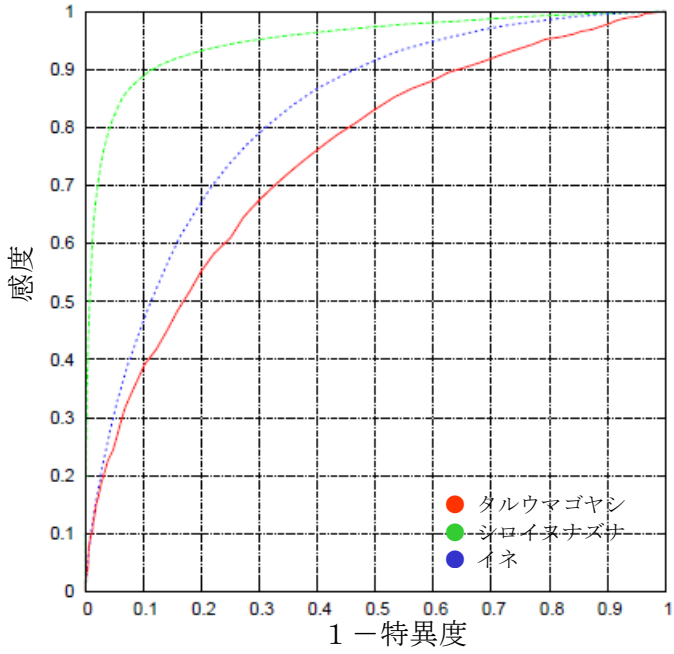

図 3.17 ３種のデータセットによる２次のマルコフモデルを使った交差検定の精度の違い。縦軸は感度、横軸は 1-特異度。

エントロピーに関しては、バックグラウンドと比較して、TSS の上流の領域[-200,0]で値が高いこと、TSS の下流の領域[0,50]で値が低いことが観察された。また、TSS でのエントロピーの減少が観察できるものの、シロイヌナズナやイネのような（図 2.13）急激な減少は観察できない。GC-skew に関しては、シロイヌナズナとイネで観察された結果と同様に TSS 周辺で増加することが確認された。2 次のマルコフモデルでの交差検定はシロイヌナズナやイネの交差検定より精度が低い結果となった。

以上の結果から、GC-skew の増加という植物での TSS 周辺の特徴を捉えてはいるものの、エントロピーの TSS での減少がシロイヌナズナとイネの２種に比べてわずかであること、２次のマルコフモデルでの交差検定の精度が２種に比べて低いこと、など考えると、先の２種のデータセットに比べるとタルウマゴヤシのデータセットは TSS を正確に捕らえられていない可能性がある。

3.4 まとめ

シロイヌナズナのデータセットについては、2次のマルコフモデルを使った予測によって、Sn＝88.66%、Sp= 89.63%、Cc= 89.14%という精度で予測が可能であった。さらに各領域のスコアに重回帰分析による重み付けをすることによって、Sn=89.38%、Sp=91.04%、Cc=73.32%という精度で予測が可能であった。シロイヌナズナに関しては K-means によるクラスタリングを行ってデータを分けても予測精度の向上がほとんどなく、プロモーター領域に高い共通性が存在すると考えられる。

イネのデータセットについては、２次のマルコフモデルを使った予測によって Sn=83.83%、Sp= 64.31%、Cc＝44.90%というシロイヌナズナに比べて低い予測精度を示した。特徴解析によってイネのプロモーター領域には多様性が存在することを確認していたため、K-means によるクラスタリングを行い、イネのデータセットを分けて、それぞれのクラスで２次のマルコフモデルを使った予測を行い予測精度が向上することを確認した。さらに各領域のスコアに重回帰分析による重み付けをすることによって、Sn=84.63%、Sp=76.55%、Cc=59.90%という精度で予測できる事を示した。

クラスタリングと２次のマルコフモデルによる予測を併用した際に、イネのデータセットを４つのクラスに分類することで、予測精度が最大値を示した。これらの各クラスにおける特徴を観測したところ、GC-skew や GC 含量の推移、Initiator などといった特徴に違いを観測した。この中の一つのクラスはシロイヌナズナにはない特徴をもつにも関わらず、全体の約４０%を占めることが確認された。イネのプロモーター領域はシロイヌナズナに比べて多様性を持っていると考えられる。

植物のプロモーター領域を一つのモデルで予測するということが可能なのかという問題については、種間で共通性の高い一部のプロモーター領域に限れば、予測が可能であった。この結果は植物の種間で共通なプロモーター領域が存在することを示唆している。予測の対象をこういった植物に共通に存在するプロモーター領域に限る事もしくはプロモーター領域の特徴によって予測モデルを複数作る事によって、植物を共通なモデルで予測可能であることが示唆された。

第４章

結論

植物のプロモーター領域（TSS 周辺）には２塩基から６塩基の短い繰り返し配列が多数存在すること事が報告されている。本研究では、シロイヌナズナとイネについて、繰り返し配列のないプロモーター領域でも、di-mer や tri-mer といった配列が高頻度に存在すること、及びこれらは TSS 周辺の GC-skew や GC 含量の推移と大きく相関があることを明らかにした。転写因子は、繰り返し配列以外にも、このような短い配列が高頻度に出現するような領域を認識している可能性がある。

di-mer や tri-mer の塩基組成については、２種に共通して、TSS 周辺に C と T が多数存在した。一方、それらの出現頻度には違いが確認できた。di-mer や tri-mer の出現パターンを学習するために、２次のマルコフモデルと重回帰分析による位置のスコアに対する重み付けの併用によって、コアプロモーター領域に限り予測を試みたところ、シロイヌナズナについては高い精度にて予測することが出来た（感度 89.38%、特異度 91.04%）。単純なモデルにも関わらず、このような精度を得たことは、プロモーター領域周辺の塩基組成が種内で共通していることを示していると考えられる。一方イネでは、コアプロモーター領域を４クラスに分け、それぞれ別にモデルを学習させることで高い予測精度を示した（感度 84.63%、特異度 76.55%）。そのうち一つのクラスは全正解セットの約 20%を占め、シロイヌナズナと類似した特徴をもつことを明らかにした。また、このクラスをシロイヌナズナの予測モデルを使って予測すると、高い予測精度を示すことが分かった。このクラスは進化的に保存されているコアプロモーター領域である可能性がある。

今後、プロモーター領域の予測精度を向上させるためには、コアプロモーター領域の特徴のみならず、繰り返しの有無、開始コドンの位置や UTR の長さなど、他の情報を組み込む必要があるだろう。また、イネのプロモーター領域がそれぞれ特徴の異なる複数のクラスに分類できることは進化の観点から興味深い。他の植物のデータを利用して、進化的に保存されたプロモーター領域を明らかにすることや、それらとその下流にある

遺伝子の関係のより詳細な解析を行うことは重要であろう。

謝辞

本研究を進めるに際しまして、暖かいご指導と研究の場を提供して下さった高木利久教授に心より感謝致します。また、直接の指導をして下さった野口英樹先生に心より感謝致します。研究室での生活セミナー等において、適切なアドバイスと励ましを頂きました小池麻子助教授、Steven Kraines 助教授、大武美保子先生、星山大介先生、牧野貴樹先生、水谷治央先生に感謝致します。小野尚孝さん、山本泰智さん、大類太郎さんには貴重なアドバイスを頂きました。高木研究室の先輩として常に適切なアドバイスをしてくださった道菅紳介さんには公私共々大変お世話になりました。また、高木研究室の学生である深川浩志さん、高橋康二さん、山口大輔さん、石井奈都さん、朴重鎬君、岩崎渉君、赤田庸平君、酒田理人君には日々の生活の中で様々な協力を頂きました。最後に、この研究に際してすべてを暖かく見守ってくれた、愛する両親に心より感謝致します。

皆様の暖かいご協力がなければ本研究を進めることはできませんでした。
このご恩は一生心に刻んで参ります。ありがとうございました。

参考文献

Altschul, SF., Gish, M., Miller, W., Myers, EW., Lipman, DJ. (1990) Basic local alignment search tool. *J Mol Biol.* **215(3)**:403-10.

Antequera, F., Bird, AP. (1988) Unmethylated CpG islands associated with genes in higher plant DNA. *EMBO J.* **7(8):** 2295-2299.

Arabidopsis Genome Initiative, (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*. **14**4: 796-815.

Bajic, VB., Seah, SH. (2003) Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.* **13(8):** 1923–1929.

Bajic, VB., Tan, SL., Suzuki, Y., Sugano, S. (2004) Promoter prediction analysis on the whole human genome. *Nat Biotechnol.* **22(11):** 1467-1473.

Bender, J. (2001) A vicious cycle: RNA silencing and DNA methylation in plants. *Cell.* **106:** 129–132.

Carlos, M., Erich, G. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*. 6-25.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., Schneider, C. (1996) Highefficiency full-length cDNA cloning by biotinylated cap trapper. *Genomics*. **137:** 327-336.

Davuluri, RV., Grosse, I., Zhang, MQ. (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet.* **29(4):** 412-417.

Down, TA., Hubbard, TJ. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12(3):** 458-461.

Feltus, FA., Lee, EK., Costello, JF., Plass, C., Vertino, PM. (2003) Predicting aberrant CpG island methylation. *Proc Natl Acad Sci USA.* **100(21):** 12253-12258.

Finnegan, EJ., Kovac, KA. (2000) Plant DNA methyltransferases. *Plant Mol Biol.* **43:** 189–201.

Florea, L., Hartzell, G., Zhang, Z., Rubin, GM. and Miller, W. (1998) A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. *Genome Res.* **8:** 967-974.

Fujimori, S., Washio, T., Tomita, M. (2005) GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics*. 6-26.

Fujimori, S., Washio, T., Higo, K., Ohtomo,Y., Murakami, K., Matsubara, K., Kawai, J., Carninci, P., Hayashizaki, Y., Kikuchi, S., Tomita, M. (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett.* **554(1-2):** 17-22.

Gardiner-Garden, M., Frommer, M. (1987) CpG islands in vertebrate genomes. *J Mol Biol.* **196(2):** 261-282.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions., A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp.japonica). *Science.* **296:** 92–100.

Green DM, Swets JA. (1966) Signal detection theory and psychophysics. *Wiley, New York.*

Jeddeloh, JA., Bender, J., Richards, EJ. (1998) The DNA methylation locus DDM1 is required for maintenance of gene silencing in Arabidopsis. *Genes Dev.* **12:** 1714–1725.

Kanhere, A., Bansal M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.* **33:** 3165-3175.

Kikuchi S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al*; Rice Full-Length cDNA Consortium; National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team; Foundation of Advancement of International Science Genome Sequencing & Analysis Group; RIKEN. (2003) Collection, mapping, and annotation of over 28 000 cDNA clones from *japonica* rice. *Science.* **301**, 376–379.

Kikuno, R., Nagase, T., Waki, M., Ohara, O. (2002) HUGE: A database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **30:** 166–168.

Kooter, JM., Matzke, MA., Meyer, P. (1999) Listening to the silent genes: transgene silencing, gene regulation and pathogen control. *Trends Plant. Sci* **4:** 340–347.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics.* **13:** 1095-1107.

Ogihara, Y., Mochida, K., Kawaura, K., Murai, K., Seki, M., Kamiya, A., Shinozaki, K., Carninci, P., Hayashizaki, Y., Shin-I, T., Kohara, Y., Yamazaki, Y. (2004) Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. *Genes Genet Syst.* **79(4):**227-232.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, DP., Bult, C., Hume, DA., Quackenbush, J., Schriml, LM., Kanapin, A., Matsuda, H., Batalov, S., Beisel, KW., Blake, JA., Bradt, D., Brusic, V., Chothia, C., Corbani, LE., Cousins, S., Dalla, E., Dragani, TA., Fletcher, CF., Forrest, A., Frazer, KS., Gaasterland, T.,

Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, IJ., Jarvis, ED., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, RM., King, BL., Konagaya, A., Kurochkin, IV., Lee, Y., Lenhard, B., Lyons, PA., Maglott, DR., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, WJ., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, JU., Qi, D., Ramachandran, S., Ravasi, T., Reed, JC., Reed, DJ., Reid, J., Ring, BZ., Ringwald, M., Sandelin, A., Schneider, C., Semple, CA., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, MS., Teasdale, RD., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, LG., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, ES., Rogers, J., Birney, E., Hayashizaki, Y.; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* **420(6915):** 563-573.

Richards, EJ., Elgin, SC. (2002) Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell.* **108:** 489-500.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., van, de, Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* **132(3):** 1162-1176.

Sakurai, T., Satou, M., Akiyama, K., Iida, K., Seki, M., Kuromori, T., Ito, T., Konagaya, A., Toyoda, T., Shinozaki, K. (2005) RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucleic Acids Res.* **33:** 657-650.

Schmid, C.D., Praz, V., Delorenzi, M., Perier, R., Bucher, P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension *Nucleic Acids Res.* **32:** D82–D85.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M.,

Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A., Shinozaki, K. (2002a) Functional annotation of a full-length Arabidopsis cDNA collection. *Science.* **296:** 141-145.

Shahmuradov, IA., Solovyev, VV., Gammerman, AJ. (2005) Plant promoter prediction with confidence estimation. *Nucleic Acids Res.* **33:** 1069-1076.

Shahmuradov, I., Gammerman, A., Hancock, J.M., Bramley, P.M., Solovyev, V.V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.* **31:** 114–117.

Shi, X., Wang, X., Li, Z., Zhu, Q., Tang, W., Ge, S., Luo, J. (2006) Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias. *Gene*. **376:** 199-206.

Suzuki, Y., Yoshimoto-Nakagawa, K., Maruyama, K., Suyama, A., Sugano, S. (1997) Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene.* **200:** 149-156.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De, Moor, B., Rouzé, P., Moreau, Y. (2001) A higher order background model improves the detection of regulatory elements by Gibbs Sampling. *Bioinformatics.* **17(12):** 1113-1122.

Yudate, HT., Suwa, M., Irie, R., Matsui, H., Nishikawa, T., Nakamura, Y., Yamaguchi, D., Peng, ZZ., Yamamoto, T., Nagai, K., Hayashi, K., Otsuki, T., Sugiyama, T., Ota, T., Suzuki, Y., Sugano, S., Isogai, T., Masuho, Y. (2001) HUNT: Launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res.* **29:** 185–188.

阿久津達也, 浅井潔, 矢田哲士. (2001) 『バイオインフォマティクスー確率モデルによる遺伝子配列解析ー』医学出版.