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
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
学籍番号　56911　巻島　健志
指導教員　高木利久　教授

論文要旨

1. 導入

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。

2. データセット

シロイヌナズナについては、cDNA の配列を RARGE から、ゲノム配列を NCBI からダウンロードした。イネについては、cDNA 配列を TIGR から、ゲノム配列を IRGSP からダウンロードした。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。また、[-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。

3. 結果と考察

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[1]。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[2]、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（図 1, 2）、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。

4. まとめ

本研究では、シロイヌナズナについてはプロモーター領域の塩基組成に共通性がみられること、イネについてはそれに多様性がみられる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。

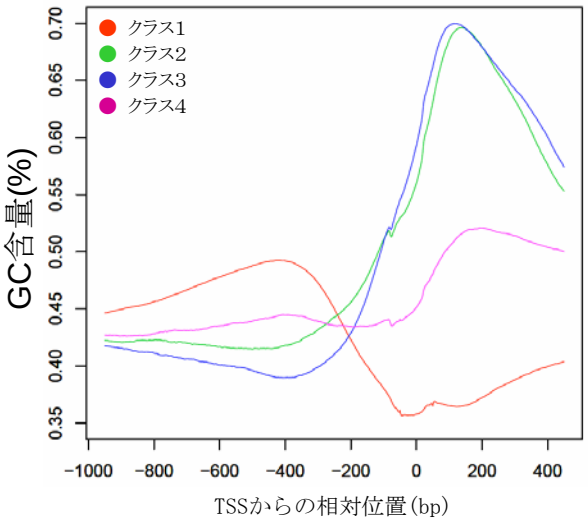

図 1 イネの各クラスでの GC 含量の推移。縦軸は GC 含量、横軸は TSS からの相対位置(bp)。

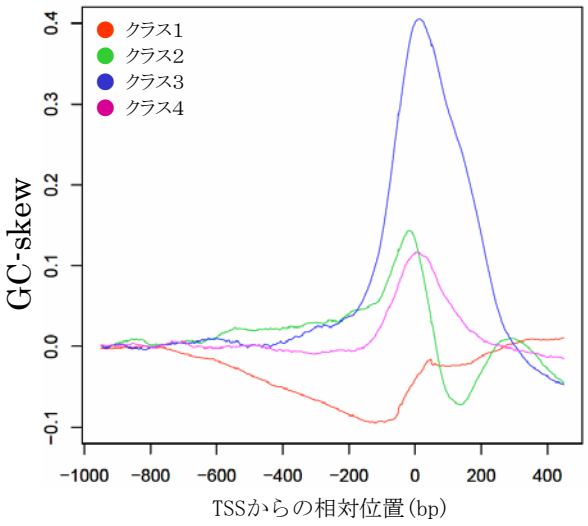

図 2 イネの各クラスでの GC-skew の推移。縦軸は GC-skew、横軸は TSS からの相対位置(bp)。

参考文献

1. Fujimori, S. *et al.* (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett.* **554(1-2):** 17-22.
2. Rombauts, S. *et al.* (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* **132(3):** 1162-1176.