助成番号　210676

５－２

研 究 成 果 報 告 書

研究テーマ（和文）	エンハンサーに関わる DNA 配列と遺伝子発現に影響する転写因子の網羅的な探索		
研究テーマ（英文）	Systematic discovery of enhancer-associated DNA sequences and transcription factors affecting the regulation of gene expression.		
研究期間	2021年 ～ 2024年	研究機関名	早稲田大学、東京電機大学、東京科学大学

研究代表者	氏名	（漢字）	大里　直樹
		（カタカナ）	オオサト　ナオキ
		（英文）	Naoki Osato
	所属機関・職名		東京科学大学・特任准教授
共同研究者 （計＿＿＿名） ＊２名をこえる場合は、【別紙追加用紙】(P3)に３人目以降を追記してください。	氏名	（漢字）	
		（カタカナ）	
		（英文）	
	所属機関・職名		
	氏名	（漢字）	
		（カタカナ）	
		（英文）	
	所属機関・職名		

概要 （600 字～800 字程度にまとめてください。）

インシュレータは、エンハンサーに結合する転写因子と遺伝子との相互作用を遮断する“仕切り”として機能する。脊椎動物では、転写因子 CTCF が主要なインシュレータとして知られているが、他の転写因子や DNA 結合タンパク質の関与については十分に解明されていない。本研究では、インシュレータに関与する転写因子および DNA 結合タンパク質を網羅的に予測する深層学習手法を開発した。

意外なことに、ヒト線維芽細胞において 97 種類の転写因子がインシュレータに関わると予測され、そのうち CTCF を含む 23 種類が、文献報告により、インシュレータ機能やクロマチン構造を介した転写制御との関連が示されていた。また、CTCF と近接する場合もあるが、MAZ は CTCF とは独立してインシュレータとして機能することが近年の研究で報告されている。

本研究では、転写因子の DNA 結合位置が、クロマチン相互作用データに基づいて同定されたインシュレータ領域に多く存在することを確認した。さらに、CTCF の結合モチーフの向きの偏り（方向性バイアス）を正確に予測できただけでなく、他の DNA 結合タンパク質の結合モチーフにも方向性の偏りが見られた。

他の研究により、MyoD や MeCP2 がクロマチンループ形成に関与していることや、CTCF およびその共働因子であるコヒーシンが欠損しても維持されるループ構造の存在が報告されており、CTCF・コヒーシン以外のループ形成メカニズムが存在する可能性が示唆されている。本研究でも、MyoD の DNA 結合位置が、骨格筋の分化と機能に関わる遺伝子の近傍のインシュレータサイトに存在することを見出した。また近年、CTCF と液−液相分離（phase separation）との関連も示されており、ループ形成が相分離を介してインシュレータ機能を発揮している可能性も示唆された。

これらの結果から、転写因子および DNA 結合タンパク質による遺伝子発現制御に、新たな役割が存在することが明らかになった。

さらに本研究の成果をもとに、ChIP-seq のような実験データを用いずに、ヒトの多様な細胞や組織で得られた遺伝子の共発現データを活用し、DNA/RNA 結合タンパク質が各細胞・組織で制御する遺伝子やその機能的役割を、深層学習により学習・予測する手法を新たに開発した。この手法により、ChIP-seq 等の実験データの取得が困難な場合でも、DNA/RNA 結合タンパク質の機能予測が可能となる。

発表文献（この研究を発表した雑誌・図書について記入してください。）

雑誌	論文課題	Systematic discovery of regulatory motifs associated with human insulator sites.					
	著者名	Naoki Osato et al.	雑誌名	bioRxiv https://doi.org/10.1101/2024.01.20.573595			
	ページ	～	発行年	2 0 2 4	巻号		
雑誌	論文課題	Functional prediction of DNA/RNA-binding proteins using deep learning based on gene expression correlations.					
	著者名	Naoki Osato	雑誌名	bioRxiv https://doi.org/10.1101/2025.03.03.641203			
	ページ	～	発行年	2 0 2 5	巻号		
図書	書名	ヒト遺伝子発現制御のインシュレータ機能に関わる DNA 結合タンパク質の解析と応用					
	著者名	大里直樹	雑誌名	情報処理学会研究報告（第 79 回バイオ情報学研究会）http://id.nii.ac.jp/1001/00238447/			
	出版社	情報処理学会	発行年	2 0 2 4	総ﾍﾟｰｼﾞ	5 ページ	
図書	書名	Topics 機械学習による遺伝子転写制御に関わる因子の探索 月刊細胞 2022 年 11 月号					
	著者名	大里直樹、浜田道昭					
	出版社	ニューサイエンス社	発行年	2 0 2 2	総ﾍﾟｰｼﾞ	3 ページ	
図書	書名	Topics 実験データ解析と機械学習による転写制御因子の探索 月刊細胞 2021 年12月号					
	著者名	大里直樹、浜田道昭					
	出版社	ニューサイエンス社	発行年	2 0 2 1	総ﾍﾟｰｼﾞ	6 ページ	

英文抄録（100 語～200 語程度にまとめてください。）
Insulator proteins function as barriers to enhancer–promoter interactions (EPIs). The primary insulator protein in vertebrates is CTCF, a DNA-binding protein (DBP). However, other DBPs associated with the insulator function of EPIs remain largely unidentified. To address this, we developed a systematic and comprehensive deep learning–based approach for identifying DNA motifs of DBPs associated with insulators.

Using this method, we identified 97 directional and minor non-directional motifs in human fibroblast cells, corresponding to 23 DBPs implicated in insulator function, including CTCF and other transcriptional regulators previously reported. The estimated orientation bias of CTCF was consistent with the directional bias observed in chromatin interaction data. These motifs were significantly more enriched at insulator sites separating repressive and active chromatin regions, at chromatin interaction boundaries, and at splice sites, compared to those of other DBPs
.
For example, we found that a MyoD-binding site—known as a key regulator in muscle differentiation—was located at an insulator site near a gene involved in skeletal muscle function. Furthermore, our insulator-pairing model suggests that both homologous and heterologous insulator–insulator pairing interactions are orientation-dependent. These findings advance our understanding of novel transcriptional regulatory mechanisms.

Building on this deep learning framework, we further developed a method to predict the regulatory target genes and functional roles of DNA/RNA-binding proteins in individual cell or tissue types, using gene co-expression data across diverse biological contexts, without relying on the experimental identification of DNA/RNA-binding sites.