

# ヒト遺伝子発現制御のインシュレータと関わる DNA 結合タンパク質の発見

大里 直樹<sup>†</sup> 浜田 道昭<sup>‡</sup>

早稲田大学 理工学術院総合研究所<sup>†</sup> 早稲田大学 先進理工学部電気・情報生命工学科<sup>‡</sup>

## はじめに

ヒトの遺伝子 (Gene) の発現制御の基本的な転写制御により、ゲノム DNA にコードされた mRNA 配列の転写が起こる。遺伝子の転写開始点近傍のプロモーター (Promoter) 領域だけでなく、遺伝子から離れた位置 (Enhancer 等) のゲノム DNA に結合する転写制御因子 (転写因子) が遺伝子の転写開始点の周辺と相互作用し、転写を制御することが明らかとなった (図 1)。

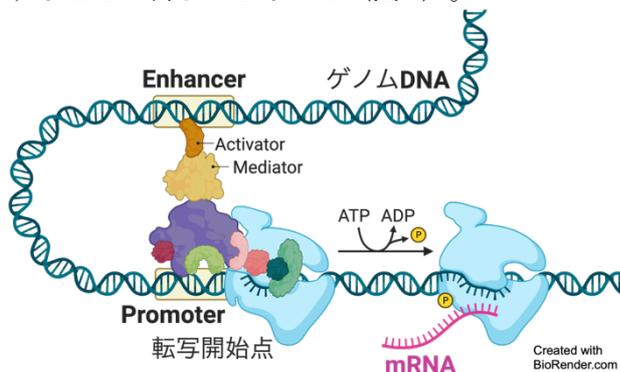


図 1 ゲノム DNA の転写制御

遺伝子から離れた位置の Enhancer 等と近傍の Promoter の相互作用の仕切り (インシュレータ) として機能する DNA 結合タンパク質が報告されている。脊椎動物では、CTCF が主要なインシュレータとして知られ、DNA のループ構造を束ねる位置に二つ (二量体) の CTCF が結合する。CTCF は Forward-Reverse の向き (逆向き) に DNA のループに結合することが多い (図 2)。

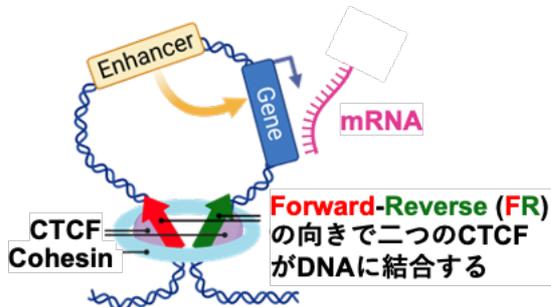


図 2 インシュレータ機能と関わる CTCF

Systematic discovery of regulatory motifs associated with human insulator sites reveals the role of directional bias

<sup>†</sup> Naoki Osato, Waseda University

<sup>‡</sup> Michiaki Hamada, Waseda University

Enhancer はループの内側の遺伝子と相互作用するが、外側の遺伝子とは相互作用しない。

本研究では CTCF 以外のインシュレータ機能と関わる DNA 結合タンパク質を網羅的に探索した。生物実験解析で多くの DNA 結合タンパク質について解析することは手間と費用がかかり困難である。

## 方法

DNA 結合タンパク質の結合位置の解析のために、DNA 結合タンパク質の DNA 結合配列とヒト線維芽細胞のオープンクロマチン領域、DNA 結合タンパク質の ChIP-seq データを用いた。遺伝子の mRNA の転写開始点の周辺の DNA 結合タンパク質の結合位置を入力データとして、細胞や組織の遺伝子発現量を学習し予測する、CNN (畳み込みニューラルネットワーク) を構築した。インシュレータ機能に関わる DNA 結合タンパク質の DNA 結合位置を仮定したときに、Enhancer や Promoter に結合する転写因子による、遺伝子発現量の予測への影響を調べるために、説明可能な AI の手法 (DeepLIFT) で Contribution score を計算し、統計検定 (Mann-Whitney *U* test) で比較した。インシュレータと関わる DNA 結合タンパク質を選択する手法を開発した。

## 結果

既知のインシュレータ機能に関わる DNA 結合タンパク質 (CTCF, Cohesin) を含め、99 種類の DNA 結合タンパク質が、インシュレータと関わることが示された。

論文検索により、インシュレータや、ゲノム DNA の相互作用や構造変化により転写制御に関わる DNA 結合タンパク質が 23 種類含まれていた。2021 年に DNA 結合タンパク質で転写因子の MAZ がインシュレータと関わるということが生物実験解析により報告された。

二つの CTCF が DNA のループに結合するときに、DNA 結合配列の向きに偏りがあることが生物実験解析により報告されており、その割合は 76.4% FR (Forward-reverse), 11.0% FF (Forward-Forward), 10.3% RR (Reverse-Reverse), 2.3% RF (Reverse-Forward) で、主に FR の向きにな

り、FR と RR が同じ程度、RF はほとんど見られない。本研究で、向きの偏りによる、遺伝子発現量の予測への影響 (Contribution score, 貢献度、重要度) を、説明可能な AI の手法 (DeepLIFT) で解析した結果、同様の傾向が示され、深層学習による予測が正確であることが示された (図3)。99 種類の DNA 結合タンパク質のうち、大部分 (89 種類) が DNA 結合配列の向きの偏りを示した。

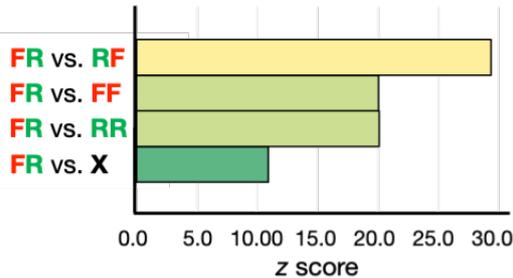


図3 CTCF の DNA 結合配列の向きの違いを比較する Contribution score の分布の統計検定 X は向きの違いを考慮しないことを示す。

mRNA が転写される活性のある領域と、転写されない抑制された領域を公共データベースの実験データを用いてゲノムブラウザ (UCSC Genome browser) で表示した。その二つの領域の境界 (インシュレータサイト) にある、99 種類の DNA 結合タンパク質の DNA 結合位置と、結合位置と重なるゲノム DNA の相互作用の実験データを可視化し、DNA 結合位置と相互作用、活性と抑制の領域の対応について確認した (図4)。

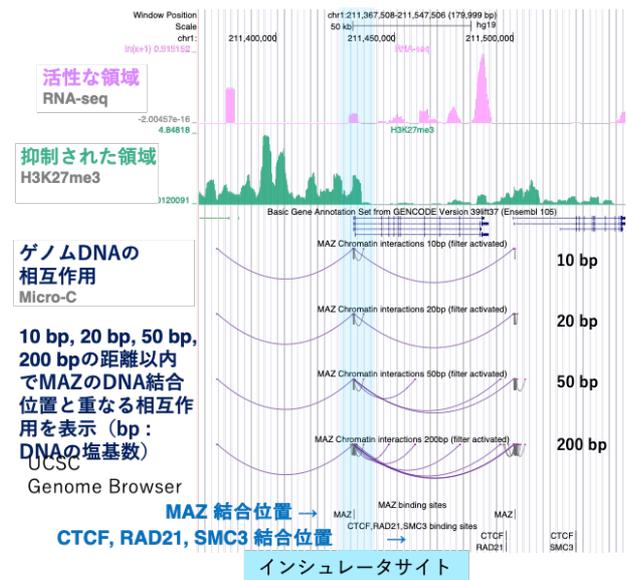


図4 インシュレータサイトと DNA 結合位置、相互作用のゲノムブラウザによる表示

その結果、インシュレータサイトにある、DNA 結合位置と重なる相互作用が、遺伝子の転写の活性と抑制の領域とよく適合することがわかった。また 99 種類の DNA 結合タンパク質は、他の DNA 結合タンパク質よりも統計的に有意に多く、インシュレータサイトに結合することが示された。これらの結果から DNA 結合タンパク質の DNA 結合がゲノム DNA の相互作用と関わり、相互作用のループにより、遺伝子の転写が活性な領域と抑制された領域が区切られることが示唆された。

考察

脊椎動物では、主に CTCF や Cohesin (RAD21, SMC3) がインシュレータ機能に関わると考えられていたが、深層学習を用いた解析から、CTCF 以外にもインシュレータサイトと関わる DNA 結合タンパク質が多く見つかった。2022 年に CTCF が相分離に関わることが生物実験解析により報告された。CTCF によるゲノム DNA のループの形成により、ループで相分離が起こり、ループの内側に転写凝集体とよばれる、転写制御に関わる因子が集まる現象が観察された。またループを束ねる位置に結合する二つの CTCF の DNA 結合配列の向きの偏りは、二つの CTCF が結合するときの構造により生じる。

これらのことから、99 種類の DNA 結合タンパク質の大部分で DNA 結合配列の向きの偏りが観察されたのは、ループの形成による構造的な制約のために生じて、ループの形成により、相分離や転写凝集体の形成が起こり、インシュレータと関わることを示唆された。インシュレータ機能は CTCF などの個々の DNA 結合タンパク質や転写因子の機能というよりも、相分離や転写凝集体の形成により生じたことが示唆された。

さらにゲノムブラウザの観察から、99 種類の DNA 結合タンパク質が、細胞や組織の種類に特異的な Alternative form の制御に関わる可能性が示唆された。細胞や組織の種類により、インシュレータサイトが変わり、遺伝子の転写される領域と重なるインシュレータサイトの違いにより、同じ遺伝子から転写される領域が異なる mRNA が生成される。組織特異的な Alternative form は転写開始点と終結点の長さの違いが多いことが報告されており、インシュレータサイトの違いが原因となり得る。

ゲノムのインシュレータサイトと関連する DNA 結合タンパク質の種類と DNA 結合位置、相互作用の情報をもとに、今後、生物実験解析によるインシュレータ機能や仕組みの全体像の解明を効率的に進めることができる。