

## 科学研究費助成事業 研究成果報告書

平成 29 年 6 月 26 日現在

機関番号：82502

研究種目：基盤研究(C) (一般)

研究期間：2014～2016

課題番号：26330339

研究課題名(和文) 計算と実験の融合によるDNA結合タンパク質の塩基配列選択能の高度化

研究課題名(英文) Improvement of sequence specificity of DNA-binding protein using computational and experimental approaches

研究代表者

河野 秀俊 (KONO, HIDETOSHI)

国立研究開発法人量子科学技術研究開発機構・関西光科学研究所 量子生命科学部・グループリーダー(定常)

研究者番号：40291918

交付決定額(研究期間全体)：(直接経費) 3,600,000円

研究成果の概要(和文)：バイオインフォマティクス、分子シミュレーション、B1Hスクリーニングを用い、2つのDNA結合ドメインを連結させた、塩基配列選択性の高いタンパク質を作成することができた。各ドメインの結合親和性を下げることで、片方のドメインのみで結合してしまうことを避け、配列選択性の向上を実現した。また、DNAの動特性を理解するため、様々なDNA配列について分子動力学計算を行った。計算結果を機械学習させ、任意のDNAの配列に対して、動特性を予測するシステムを構築した。これを用いてゲノム配列を解析し、転写因子の結合領域のみならず前後200bpの領域において、DNAの動特性に特徴があることを見出した。

研究成果の概要(英文)：We successfully created a designed-protein which can bind to a target DNA sequence with high specificity. In the design, we connected two DNA-binding proteins after engineered one mutation which reduces the affinity of the monomer. Then, we combined these domains to create a dimer protein and confirmed that it binds to the target sequence by B1H assay. In addition, to understand the DNA sequence dependent, physico-chemical property, we carried out molecular dynamics simulations on many DNA sequences. We extracted some feature parameters from the sampled conformations using a machine learning technique. We found that the values of the feature parameters at transcription binding sites as well as within 200bp up and down streams from the binding sites are different from those of non-binding sites.

研究分野：生命情報科学、生物物理学、蛋白質科学

キーワード：DNA結合蛋白質 分子シミュレーション 生体生命情報学 バイオインフォマティクス 蛋白質設計 B1H

## 1. 研究開始当初の背景

遺伝情報の発現は、タンパク質が特定の塩基配列に結合することによって精密に制御されている。この結合がおかしくなると、転写、発現、分化といった生命現象の根幹に異常をきたし、がんなど様々な疾病を引き起こしてしまう。近年、ゲノムの塩基配列情報にはよらない、ヒストンタンパク質の化学修飾や DNA のメチル化などによる遺伝学 (エピジェネティクス) も盛んに研究され、今やエピゲノム創薬という言葉までできている。しかし、ジェネティクスにおいてもエピジェネティクスにおいても、これらの現象の根本はタンパク質と DNA の塩基配列特異的な相互作用であることには変わりはない。転写因子などの DNA 結合タンパク質がどの遺伝子を制御しているかを調べるために、免疫沈降法によるゲノムワイドな結合配列解析やマイクロアレイ解析が行われてきた。これらの解析では、遺伝子と DNA 結合タンパク質の関係を明らかにしてくれるものの、分子認識機構まではわからない。そのため、タンパク質と DNA の複合体の立体構造解析が行われ、2013 年 10 月末の時点で 4428 件の立体構造が決定されている。申請者らは、これらの複合体データを解析し、アミノ酸残基と塩基との相互作用に明らかな傾向はあるものの、1 対 1 の関係はなく、同じアミノ酸であっても前後の配列によっては異なる塩基を認識する冗長な関係になっていることを見出した。さらに、DNA 結合タンパク質が結合しやすい (つまり、認識する) 塩基配列を予測する方法を開発した。また、想定した主鎖構造に対して最適なアミノ酸配列を設計する方法を開発[4]し、その方法を用いて膜タンパク質の水溶性タンパク質の改変に世界に先駆けて実現した。さらに、酸化還元能を持つ新規タンパク質の創製に成功した。2011 年には DNA 結合タンパク質の認識塩基配列の予測方法とアミノ酸配列設計方法を組み合わせることによって、野生型よりも配列選択性の高い亜鉛フィンガー型 DNA 結合タンパク質の創製に成功している。また、詳細な自由エネルギー計算によって、ラックリプレッサータンパク質の DNA 特異的結合と非特異的結合の違いを調べ、特異的結合の場合には結合自由エネルギーが鋭く低くなるのに対し、非特異的結合ではそのような鋭いエネルギー

の減少は見られないことなど、結合状態の違いを分子論的に明らかにしている。

このような研究を通して、申請者は DNA 結合タンパク質を新たな分子マーカーや遺伝子治療、ゲノム創薬に応用できるのではないかと考え、本研究の着想に至った。野生型の DNA 結合タンパク質を、そのまま遺伝子治療やゲノム創薬に応用するには塩基配列選択性や結合強度が低い。実際、同じ転写因子が複数の遺伝子を制御し、結合する塩基配列もひとつに限らず一群の似た配列に結合することが知られている。そのため、既知の転写因子をそのまま薬に用いると、意図しないところで働き、重篤な影響を起しかねないので、転写因子の塩基配列特異性の向上が望まれている。また、次世代シーケンサーの出現により、大量の転写因子結合データが蓄積されているものの、DNA やタンパク質の構造まで考慮に入れて解析されてはならず、構造や動的な観点を導入することで、転写因子の結合メカニズムが理解できると考えられる。

## 2. 研究の目的

本研究では、申請者がこれまで培ってきたバイオインフォマティクス、分子シミュレーション、タンパク質設計の研究を統合することにより、野生型よりもより塩基配列選択性の高いタンパク質や野生型とは異なる塩基配列に結合するタンパク質を設計することを目指す。同時に、次世代シーケンサーによって得られた転写因子結合データを、DNA の配列依存的な構造及び動的な観点から評価し、結合メカニズムを理解する。

## 3. 研究の方法

### (1) エングレイルドホメオドメインタンパク質の配列特異性の向上

本研究では、2 つのエングレイルドホメオドメインタンパク質をリンカーでつなぐこと、および、リンカーで繋いだ蛋白質に部位特異的変異を導入することを組み合わせ、配列特異性の向上を試みる。

そのために、ホモロジーモデリング法によって結合構造の推定を行い、リンカーや部位特異的変異体を設計する。設計した蛋白質は大腸菌で発現し、アフィニティ担体を用いて精製する。得られた蛋白質の結合強度は *in*

*in vitro* のゲルシフトアッセイ実験を用いて検討する。更に、*in vivo* での配列特異性を検討するために、B1H (Bacterial One-Hybrid) 法(レポーター遺伝子の発現の有無を利用して、ランダム DNA 配列を有するプラスミドライブラリから、DNA 結合タンパク質が結合したプラスミドだけを選択する方法)によって設計通りの結合特性を持つか調べる(図 1)。

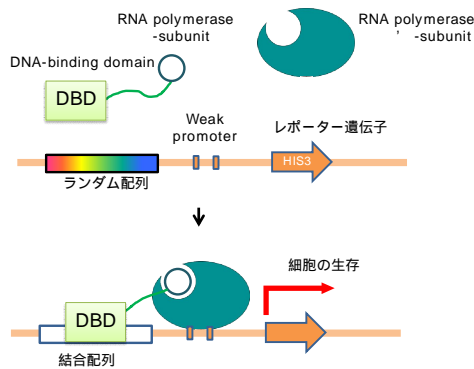


図 1 . B1H 法概要

( 2 ) 分子動力学計算から DNA 配列の動特性の抽出

12 塩基対の DNA、GCGC-NNNN-GCGC (N は、A,T,G,C のいずれか) の配列パターンは、対称性を考慮すると 136 種類ある。その 136 種類について、100 ナノ秒の分子動力学計算(MD)を行い、各配列に対する構造集団を作成する。塩基対パラメータ( 6 つ)及び塩基対ステップパラメータ( 6 つ)を用いて、構造集団の特徴解析を行う。

( 3 ) 転写因子結合部位及びその周辺領域の DNA 配列の特徴解析

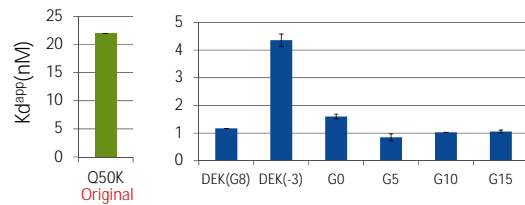
( 2 ) で得られた特徴量を用いて、転写因子 1312 種類の結合配列を評価する。

#### 4 . 研究成果

( 1 ) エングレイルドホメオドメインタンパク質の配列特異性の向上

ホモロジーモデリングによって、どのようなリンカーを用いれば 2 つのドメインが DNA にタンデムに結合しうるのかをコンピュータ上で推定した。結果、リンカーとして 0 ~ 5 残基程度のフレキシブルリンカーを用いることで、ドメイン間に立体障害を生じることなく目的とする 12bp 長のターゲット配列との結合ができることが分かった。

リンカーの長さを種々に変更して *in vitro* でのアッセイを行った結果、5bp の Gly を繋いだ G5 リンカーが目的とする 12bp 長のターゲット配列との結合強度が最も良好であることがわかった( 図 2 )。



Target sequence  
O50K: d(CGCAGTG-TAATCC-CCTCGAC)  
Others: d(CGCAGTG-TAATCC-TAATCC-CCTCGAC)

図 2. 得られた解離定数

ところが、B1H 法で目的とする配列との結合特異性を調べたところ、片方の DNA 結合ドメインの結合のみでも、B1H アッセイ系が動いてしまい、期待した認識特異性の向上は観察されなかった。ただ単に 2 つの蛋白質を繋いただけでは結合強度が必要以上に強くなってしまい、特異性の向上には繋がらなかったためと考えられる。そこで、DNA との認識配列には影響せず、しかし、DNA との結合親和性を弱める変異部位を立体構造から推測した。推定した部位に変異を入れた変異体のうち、結合親和性が 10 倍以上下がったものを選んでタンデム型につなぐことで、目的通り、倍の長さの DNA を認識するタンパク質を作成することができた(図 3)。

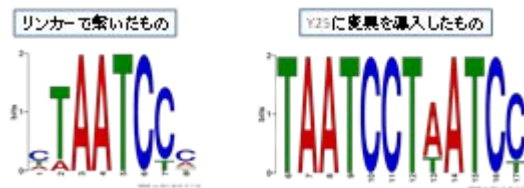


図 3 . B1H 法で得られた認識配列

( 2 ) 分子動力学計算から DNA 配列の動特性の抽出

これまでの研究では、配列に依存した DNA の平均構造は考慮されていたが、タンパク質との結合において重要である、構造の柔らかさ、つまり、構造変形のしやすさは考慮されていなかった。そこで、塩基対パラメータ及び塩基対ステップパラメータの分布状態を調べることで、構造変形のしやすさを解析した。解析に、機械学習のアルゴリズムを導入するため、各パラメータの分布を 5 つの区分に分けて、各区分の分布の偏りを配列ごとに

学習させ、構造変形のしやすさを表現した。入力を DNA の配列、出力を DNA の構造という形で学習を行った。入力 DNA 長を 1 から N まで変えて、最適な DNA 長を探した。結果、5 塩基対長が最もよく MD データを再現することが分かった(図 4)。また、各パラメータ間の相関を調べることで、5 塩基長 DNA 配列の、平均構造と構造変形能における類似性を明らかにすることができた(図 5)。

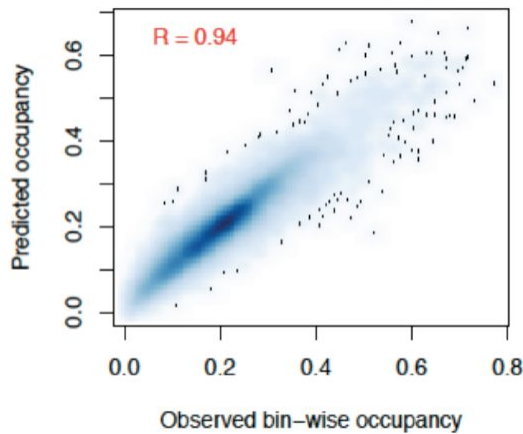


図 4 機械学習によって予測された DNA のパラメータと学習データに用いたパラメータの相関。

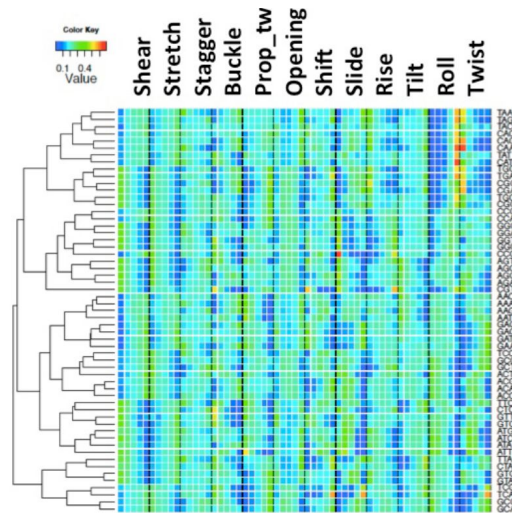


図 5 . DNA 配列と各パラメータの類似性。同じ色は、同程度の分布を持つことを表す。

### (3) 転写因子結合部位及びその周辺領域の DNA 配列の特徴解析

DNA 配列に依存した動特性の観点から、DNA 結合タンパク質の認識塩基配列を解析した。(2)の機械学習で作った予測方法を用いて、1312のDNA結合タンパク質のDNA結合部位データを解析した。転写因子が結合

する部位のみならず、その上流及び下流の領域について、動構造パラメータの平均からの「ずれ」を調べた。「ずれ」を評価することで、結合領域とそうでない領域を見分けることができる。この動特性の「ずれ」は、DNA結合領域のみならず、その上流及び下流 200 bp まで、DNA が結合しない領域と有意に違いがあることが分かった(図 6)。このことは、技術的には、広い範囲の DNA の動特性を調べることで、転写因子の結合部位の予測精度が向上することを示唆している。一方、生物学的には、転写因子が結合する部位は、転写因子が結合できるように、あらかじめ DNA がアクセスしやすいようにゲノムが構成されていることを想像させる。さらに、「ずれ」を生み出した要因を詳細に解析すると、塩基対の開きを表すパラメータが最も予測に寄与していることが分かった(図 7)。転写が起こるためには、2 重らせんは開裂しなくてはならず、そのことを考えると、妥当な結果であるように思える。

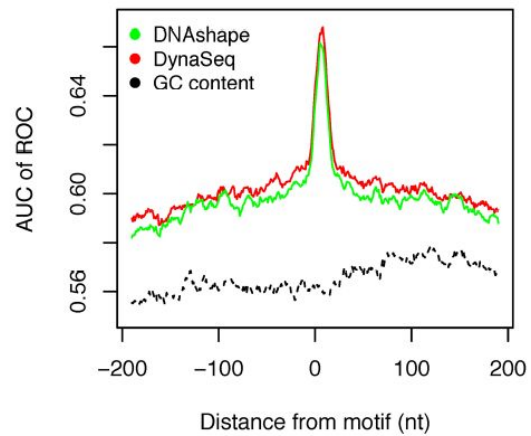


図 6 . 転写因子結合部位及び上流、下流 200 bp の動特性の「ずれ」。本研究の方法：赤丸；DNASHape 法：緑；GC 含有。



### Best predictor counts (%)

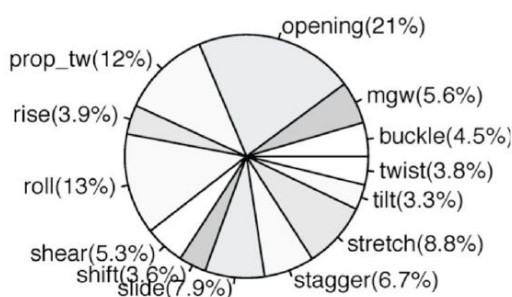


図7. 1312の転写因子に対して、結合部位とそうでない領域で最も「ズレ」が大きかったパラメータの割合。

### 5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文](計1件)

Andrabi, M.; Hutchins, A. P.;

Miranda-Saavedra, D.; Kono, H.; Nussinov, R.; Mizuguchi, K.; Ahamd, S., Predicting conformational ensembles and genome-wide Transcription Factor binding activities from DNA sequences. Scientific Reports, 7, 4071, 査読有, 2017, DOI: 10.1038/s41598-017-03199-6

[学会発表](計7件)

Sunanmi, T., Chatake, T. and Kono, H., Revisiting DNA conformations in crystal structures, 第54回日本生物物理学会年会, つくば国際会議場(茨城県つくば市), 2016年11月25-27日

河野 秀俊, 構造多形を生体高分子の全原子モデルの構築, JSBi 関西地域部会 第21回バイオメディカル研究会, グランフロント大阪(大阪府大阪市), 2016年9月6日

角南 智子, 河野 秀俊, エングレイルドホメオドメインを用いた新規転写因子設計, 第16回日本蛋白質科学会年会, 福岡国際会議場(福岡県福岡市), 2016年6月7-9日

河野 秀俊, 転写因子のアクティビティは結合サイト及びその周辺領域の DNA

特性の関係できまる, 第16回日本蛋白質科学会年会, 福岡国際会議場(福岡県福岡市), 2016年6月7-9日

角南 智子, 河野 秀俊, Designing a new artificial transcription factor based on engrailed homeodomain, 第38回日本分子生物学会, 神戸ポートアイランド(兵庫県神戸市), 2015年12月1-4日

角南 智子, 河野 秀俊, Designing a new artificial transcription factor based on engrailed homeodomain, 第15回日本蛋白質科学会年会, あわぎんホール(徳島県徳島市), 2015年6月24-26日

角南 智子, 河野 秀俊, Designing a new artificial transcription factor based on engrailed homeodomain, 第51回生物物理学会年会, 札幌コンベンションセンター(北海道札幌市), 2014年9月25-27日

### 6. 研究組織

(1)研究代表者

河野 秀俊(KONO, Hidetoshi)

国立研究開発法人量子科学技術研究開発機構・関西光科学研究所 量子生命科学研究部・グループリーダー  
研究者番号: 40291918

(2)連携研究者

角南 智子(SUNAMI, Tomoko)

国立研究開発法人量子科学技術研究開発機構・関西光科学研究所 量子生命科学研究部・主幹研究員  
研究者番号: 50554648