2007－BIO－10 (6)
2007／9／13

アミノ酸残基の空間的出現確率による DNA Binding Protein の機能部位解析

坂辻 侑華子, 沖原 伶佳, 山登 一郎, 宮崎 智
東京理科大学大学院 薬学研究科 生命情報科学研究室

遺伝子発現制御のメカニズムを解明するには、DNA Binding Protein(DBP)の結合部位の構造特性を解明する必要がある。現在の DBP の機能分類は、配列モチーフが主たる根拠となっている。しかし、配列モチーフには、DBP の構造維持に係わるものと DNA 結合の特徴を示すものが混在しており、必ずしも DNA 結合部位に着目した分類とはなっていない。一方、我々は、局所構造情報に基づいた機能部位予測を目的に開発された FCANAL を応用し、DNA 結合部位におけるアミノ酸残基配置の空間的出現確率が、構造的特徴によく対応することを見出してきた。本研究では、これを応用し、DNA 結合部位の構造特性による DBP の分類を試みた。

Analysis of functional site of DNA Binding Proteins based on spatial existing probability of amino acids

Yukako Sakatsuji, Reika Okihara, Ichiro Yamato, Satoru Miyazaki
Department of Pharmaceutical research, Tokyo University of Science

To explore the mechanism of gene regulation, it is necessary that structural characteristic of binding site on DNA Binding Proteins(DBP) which cognize the cis-element. Database of the DBPs classifies based on sequence motif. However, sequence motif is consisted of residuew related to structure maintenance and DNA binding. On the other hand, we found that spatial existing probability of amino acids in DNA Binding site coped for characteristic of structure , using FCANAL developed for identification functional site based on local structure information and. In this study, We attempt to clarify the structural characteristic of DNA binding sites.

1. はじめに

転写とは、DNA1 本鎖を鋳型に mRNA を合成する反応を示す。この転写反応は、いくつかの段階に分けることができる。RNA ポリメラーゼがプロモーターに結合して転写が開始されるまで、リン酸ジエステル結合時期、伸長、終結の 4 段階ある。上記の転写開始前時期に DNA と分子間相互作用を起こして各段階におけるステップを調節するタンパク質が DNA Binding Protein (DBP)であり、結合する DNA 部位をシスエレメント配列という。DBP の構造は様々であり、その結合部位の構造特性がシスエレメント配列認識や遺伝子発現制御に関与することにより、遺伝情報が適量発現されるように制御されている。したがって、遺伝子発現を制御する DBP のシスエレメントを認識する部位の構造特性を解明することは、遺伝子発現制御メカニズムの謎を解く上で必要不可欠である。DBP の構造特性を探索するには、機能毎に分類された DBP 立体構造データが必要となる。しかし、現在の DBP の分類はシスエレメント情報[1]や、DBP の配列モチーフに基づいたもとであり、DNA 結合部位の立体構造特性に特化した分類は行われていない。そこで本研究では、DNA 結合部位の局所立体構造情報に基づいて機能部位を特徴付けることを目的で開発してきた FCANAL (Fast Calculable Protein Function Analyzer)[2,3]を応用し、DNA 結合部位におけるアミノ酸残基の空間的出現確率から局所的な立体構造を表現し、機能との関連性を検討した。その上、DNA 結合部位の構造に基づいた DBP 分類手法の確立を試みた。

2. 配列モチーフを基点とした DBP の機能部位予測

FCANAL は、既存の機能部位に対して、アミノ酸残基の空間的出現確率による数量化（学習）を行い、機能と対応づけるプログラムである。その為、機能部位既知の立体構造情報が必要となる。また、タンパク質の機能に関連するアミノ酸残基を含むとされる配列モチーフの情報を網羅することで、機能部位情報を特徴づけれる可能性が高い。そこで我々は、FCANAL を用いて配列モチーフ周辺の局所構造情報に基づいた機能部位の特徴付けを行った。以下に Zinc Finger C2H2 (ZFC2H2)をターゲットにFCANALの概要を説明する。

2.1 データの取得

DBP の立体構造を Protein Data Bank (PDB)[4]から収集し、PROSITE[5]に登録されている配列モチーフに基づいて機能分類を行った。その結果 14 データセット作成した。その内、ZFC2H2 の配列モチーフを持つ立体構造は 26 個存在した。

2.2 局所構造のスコア化(機能部位の学習)

2.2.1 Key Residue 選出

機能部位に共通し機能の発現に最も重要と思われるアミノ酸残基を Key Residue とした。ZFC2H2 の立体構造は、α-helix、β-sheet から形成されている。このうち、機能の発現に重要なのは、α-helix である事が実験結果から明らかになっている[6]。そこで、α-helix を構成するアミノ酸残基の中で、最も高度に保存されている His を Key Residue と定めた。

<ZFC2H2 の配列モチーフ>

C- x (2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

↑Key Residue

2.2.2 局所構造定義づけ

Key Residue から半径 7～15Åの領域を局所構造と設定した。ZFC2H2 では、His から 14Åの領域を局所構造とした。その中でも特に、配列モチーフ周辺で形成された局所構造を機能部位とした。

2.2.3 局所構造におけるアミノ酸残基の距離情報と出現頻度情報

局所構造情報を用いて、アミノ酸残基の距離と出現頻度を算出した。Key Residue から局所構造を構成する全てのアミノ酸残基(aa)の距離(x)を測定した上で、Key Residue を中心に 0.5Å毎に局所構造を区切った。ZFC2H2 の局所構造の場合、29 階級に区切った。次に、機能部位(site)、機能部位以外の局所構造(bg)における各階級(c)に存在する各アミノ酸残基(aa)の個数($F^{site}(aa,c)$、$F^{bg}(aa,c)$)を数え、次式によりアミノ酸残基が階級(c)に属する確率、出現頻度($P^{site}(aa,c)$、$P^{bg}(aa,c)$)を算出した。

$$P^{site}(aa,c) = \frac{F^{site}(aa,c)}{\sum_{aa=1}^{20}\sum_{c=1}^{24} F^{site}(aa,c)}$$

$$P^{bg}(aa,c) = \frac{F^{bg}(aa,c)}{\sum_{aa=1}^{20}\sum_{c=1}^{24} F^{bg}(aa,c)}$$

2.2.4 Score Matrix の作成

機能部位および局所構造から抽出した出現頻度情報から、機能部位におけるアミノ酸残基の存在の存在確率(スコア、S(aa,c))を算出した。

$$S(aa,c) = \ln\left(\frac{P^{site}(aa,c)}{P^{bg}(aa,c)}\right)$$

得られた存在確率を用いて Score Matrix を作成した。ZFC2H2 の Score Matrix は、29 階級における 20 種類のアミノ酸残基存在確率を算出している為、580 のスコア要素を含んでいることになる。スコア要素から、機能部位の構成に重要なアミノ酸残基の特定を行った。

2.3 判別値の最適化

Score Matrix の機能部位予測精度を検証した。そこで、全ての局所構造における Score Matrix を階級毎に足し合わせた関数 DB 関数を算出した。

$$DB(x) = f(S(aa,c))$$

得られた DB 関数から、階級スコアを距離スコアに換算する為、隣接する階級の頂点を結んだスコア関数を TS 関数とした。

$$TS = \sum_{x=1}^{Num} DB(x)$$

更に、機能部位及び局所構造における独立性、そして作成したスコア行列の予測精度及び網羅性を評価する為に閾値を設定した。機能部位と予測した局所構造の

Table1. *配列モチーフによる機能部位予測結果*

F値は機能部位の予測精度を表しており、その値が2.00であれば、FCANALによる機能部位予測の精度が非常に良いことを示す。また、A値が高い程、機能部位と局所構造のTSが二極化する。

Target	Cutoff(Å)	Key residue	F	A
Homeo Box	14	Arg	1.97	4.85
ZFC2H2	14	His	2.00	5.18
POU	13	Gly	2.00	8.12
HTH_LACI	11	Ala	1.83	4.05
ZincFingerC6	14	Cys	1.93	3.21
GATA ZF	14	Cys	1.84	5.23
b ZIP	10	Arg	1.92	2.34
Fork HEAD	13	His	2.00	20.67
HTH CRP	12	Arg	1.88	2.19
TF2D	14	Tyr	1.96	5.33
MARDS BOX	14	Thr	1.98	6.37
ZF FPG	14	Cys	1.86	4.11
TF2B	14	Gly	1.93	4.62
HSF	13	ARG	1.75	3.73

TS値(TS_{site})と、機能部位以外と予測した局所構造のTS値(TS_{bg})の平均値の差をA値とし、これを閾値と設定とした。

$$A = \bar{TS}_{site} - \bar{TS}_{bg}$$

作成したスコア行列の予測精度及び網羅性を評価する為に、Pr値とPq値を用いてF値を算出した。Pr値は任意の局所構造が正しく予測された割合を示し、Pq値は局所構造群中の正しい局所構造の割合を示す。

$$F = \max_{i,j} (\Pr + Pq)$$

$$\Pr ecision(\Pr) = truehits/(truehits + falsepositives)$$

$$Recall(Pq) = truehits/(truehits + falsenegatives)$$

2.4 結果

ZFC2H2、HTH_LACIのScore MatrixとScores DataをTable1およびFigure1に記す。Scores Dataの横軸は機能部位におけるTS値を示している。ZFC2H2において、Scores DataからScore Matrixで機能部位を特徴づけることができた。しかしながら、HTHのScores Dataは機能部位と局所構造のTSが二極化せず、Score Matrixで機能部位を特徴づけることはできなかった。特徴付けに成功したデータセットはZFC2H2を含め、3データセットであった。それに対し、HTH_LACIを含めた11データセットにおいてFCANALは機能部位を特徴付けることはできなかった。

2.5 考察

FCANALは、既存の機能部位の局所構造について、活性中心となるアミノ酸残基を1つ選定することから始め、機能部位の数量化(学習)を行う。当初我々は、FCANALを用いて酵素の機能部位の数量化を試みた。その結果、機能部位の容積が比較的小さい場合に、高い予測精度を実現することが分かった。しかし、DBPについては、FCANALの精度が非常に悪いものも含まれていた。酵素におけるFCANALの予測精度の良さを考えると、DBPのDNA結合部位の局所構造によって、DNA認識のメカニズムを特徴づけることが難しいと結論づけているように見える。あるいはその原因として、学習データ不備の可能性が考えられる。そこで学習データの精度についての検証を行うことにした。

3. 学習データの検証

前述の検証ではPROSITEに登録されている配列モチーフ情報を機能部位の指標として用いたが、この配列モチーフとDNA結合部位の空間的位置関係を検証する。

3.1 データセット作成

PDBからProtein-DNA complexの立体構造データを全て抽出した(1300エントリー)。更に、DNAから3.1

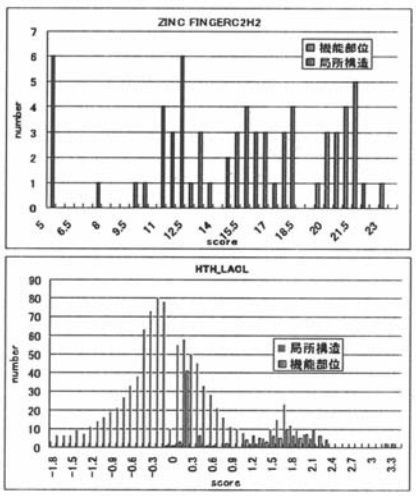

Figure1. *ZFC2H2 の Scores Data（上）と HTH_LACI の Scores Data(下)*

縦軸：局所構造数　横軸：各局所構造のTSを示す

機能部位とその他の局所構造の2つのスコアが二極化されていれば、Table1におけるA値は大きくなり、ScoreMatrixで特徴付けた機能部位はZFC2H2の機能部位において特徴できであると言える。

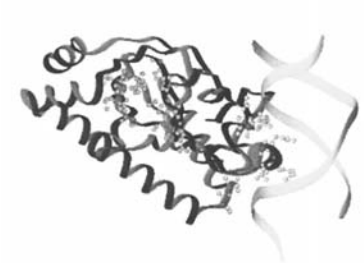

Figure2.　HTH_CRP における配列モチーフと DNA 結合アミノ酸残基の差異
配列モチーフを構成するアミノ酸残基(orange-ball)DNA に結合しているアミノ酸残基 (pink-ball)

Å以内に存在するアミノ酸残基を全て抽出し、Amino acid Binding to DNA(ABD)データベースを作成した。

3.2　結果と考察

配列モチーフに含まれるアミノ酸残基と ABD データベースのアミノ酸残基の空間的位置関係を検証した。その結果、前述の方法で機能部位をうまく特徴付けることが出来ていた DBP は、配列モチーフと DNA に結合しているアミノ酸残基がほぼ一致している傾向にあった。一方、機能部位を特徴付けれなかった DBP は、配列モチーフと DNA に結合しているアミノ酸残基が離れていた(Figure2)。すなわち、配列モチーフは必ずしも DNA 結合部位となっていないことが判明した。また、配列モチーフには DBP の立体構造形成に関与するものと、DNA 結合部位の特徴を示すものが混在していることが判っており、配列モチーフによる DBP 結合部位の特徴づけには細心の注意が必要であるといえる。

4　DNA 結合部位の機能部位予測

前検証結果より、PROSITE に登録されている配列モチーフのアノテーション情報の利用には注意が必要である。そこで、配列モチーフに依存しないタンパク質の機能情報から DBP 立体構造の機能分類を提案する。更に、DNA 結合部位の情報を FCANAL を用いて、機能部位に特徴的なアミノ酸を検討した。

4.1　データセット作成

タンパク質の機能や構造に関するデータベースとの相互参照情報を重視した UniProt[7] と ABD データベースを利用し、DBP を機能毎に分類し、FCANAL 実行用の 40 データセット作成した。

4.2　局所構造のスコア化(機能部位の学習)

4.2.1　Key Residue 選出

ABD データベースを利用し、DNA に結合しているアミノ酸残基の重心に最も近いアミノ酸残基を Key Residue とし、FCANAL を実行した。

4.3　結果

DNA 結合部位に関して機能部位予測した結果の 1 部を Table2 に示す。HTH_LACI は、大腸菌内に存在するラクトースオペロンリプレッサーである(Figure3.B)。このリプレッサーは、機能部位を介して RNA ポリメラーゼのプロモーターへの結合を阻止し、転写を抑制する。そして、ラクトース代謝酵素の発現を調節する(Figure3.A)[8,9]。FCANAL を用いて、HTH_LACI の Scores Data (Figure4)の機能部位と局所構造を二極化に成功し、機能部位を特徴づけれたと言える。

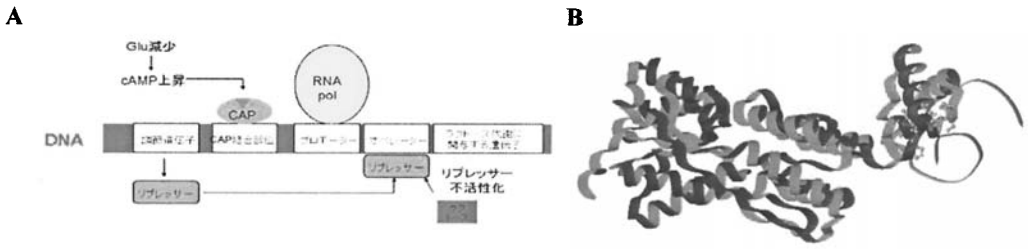

Figure3. HTH_LACI による転写調節機構(A)と立体構造(B)
(A)　FCANAL で予測した機能部位がオペレーター部位に対して機能を発現すると思われる
(B)　ピンクの球が FCANAL をもって機能部位を構成すると断定したアミノ酸残基(PDBID:1bdh)

Table2.　DNA 結合部位による機能部位予測結果

Target	Cutoff(Å)	Key residue	F	A
HTH_LACI	10	Tyr	2.00	10.23
MTAL2_YEAST	12	Arg	2.00	6.35
MTH1_HAEHA	13	Tyr	2.00	3.16
MTRB_BACST	14	Glu	2.00	7.68
MUTS_ECOLI	14	Lys	2.00	3.98
MYB_AVIMB	14	Glu	2.00	4.22
NFAC2_HUMAN	14	Tyr	2.00	5.89
NFKB1_HUMAN	14	Arg	2.00	8.12
OGG1_HUMAN	14	Val	2.00	5.40
PO2F1_HUMAN	14	Ser	2.00	7.24
PPO1_PHYPO	14	Gln	2.00	11.61
RGRO_BP434	10	Ile	2.00	5.58
RECR_BPP1	14	Gly	2.00	6.58
RDRP_BPPH6	14	Asn	2.00	2.93
ROA1_HUMAN	14	Lys	2.00	6.67
RUNX1_HUMAN	13	Val	2.00	3.32
RXRA_HUMAN	10	Arg	2.00	8.90
SRY_HUMAN	14	Phe	2.00	5.63

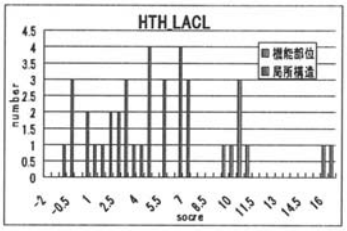

Figure4.　HTH_LACI の Scores Data

4.4　考察

全ての F 値が 2.00 となり、*2．配列モチーフによる機能部位予測*時よりも DNA 結合部位に着目して機能部位予測を行った時の方が大幅に予測精度が向上した。従って、DBP の結合部位におけるアミノ酸残基の空間的出現確率によって局所的な立体構造を表現し、機能と機能部位の対応付けが完了した。

5.　DBP の分類

前項において、DBP の機能と機能部位の関連性を見出し、新たな DBP の機能分類の手法が確立したと言える。しかし、この分類は、あくまで機能毎における分類であり、機能部位の共通性による分類に至っていない。そこで、本項では機能部位におけるアミノ酸残基の空間位置が高度に保存されている事に着目し、機能部位間の共通性を探索した。

5.1　DBP の機能部位の階層的クラスタリング

DBP の機能部位を網羅的に比較する為に、FCANAL 実行から得られた DB 関数(Figure5)の類似性をもとにユークリド距離・ウォード法[10]を用いて階層的クラスタリングを行った。これより、機能部位の空間的位置関係が類似している DBP を知ることが出来る。2 種類の DBP に対応する 2 つの DB 関数について、それぞれを DB 値 29 ポイントと隣接する階級間の傾き 29 ポイントの 58 次元ベクトルに変換し、DBP の機能部位の違いをユークリッド距離で評価する。DBP 全ての組み合わせについて、ユークリッド距離を計算し、距離行列を得た。ユークリッド距離行列の中から最も類似性が高い 2 つの DBP を合併し、1 つのクラスタを作る。そして、次々とクラスタを結合し、最終的に全ての DBP が 1 つのクラスタに合併されるまで繰り返すことで階層的構造を作成する。

DB 関数表 a と DB 関数表 b 間のユークリッド距離 D は以下の式により与える。

$$D(a,b) = \sqrt{\sum_{i=1}^{5} (a_i - b_i)^2}$$

5.2　結果・考察

DBP の DB 関数表 40 レコードを用いてクラスタリングを行った結果を Figure7.A に示す。ZINC FINGER C2H2、AREA、GATA は Zinc Finger type の立体構造 Domain、また PurR、HTH_LACI、HTH_CRP は HTH type の立体構造 Domain を持つ。しかし、得られたデンドログラムは、それらが近隣にクラスタリングされておらず、DBP の機能部位と立体構造 Domain に相関がないように思われる。あるいは、機能部位を構成するアミノ酸とは無関係に、偶然各階級における DB 関数の傾向が似通っていた可能性も考えられる。

そこで、我々は分離能力が比較的良いクラスタに着目し、DBP を 1 つのデータセットに集め、再度 FCANAL を実行した。例えば、クラスタ 1(Figure7.B)では

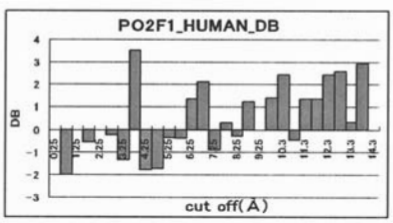

Figure5.　DBPPO2F1_HUMAN の DB 関数表

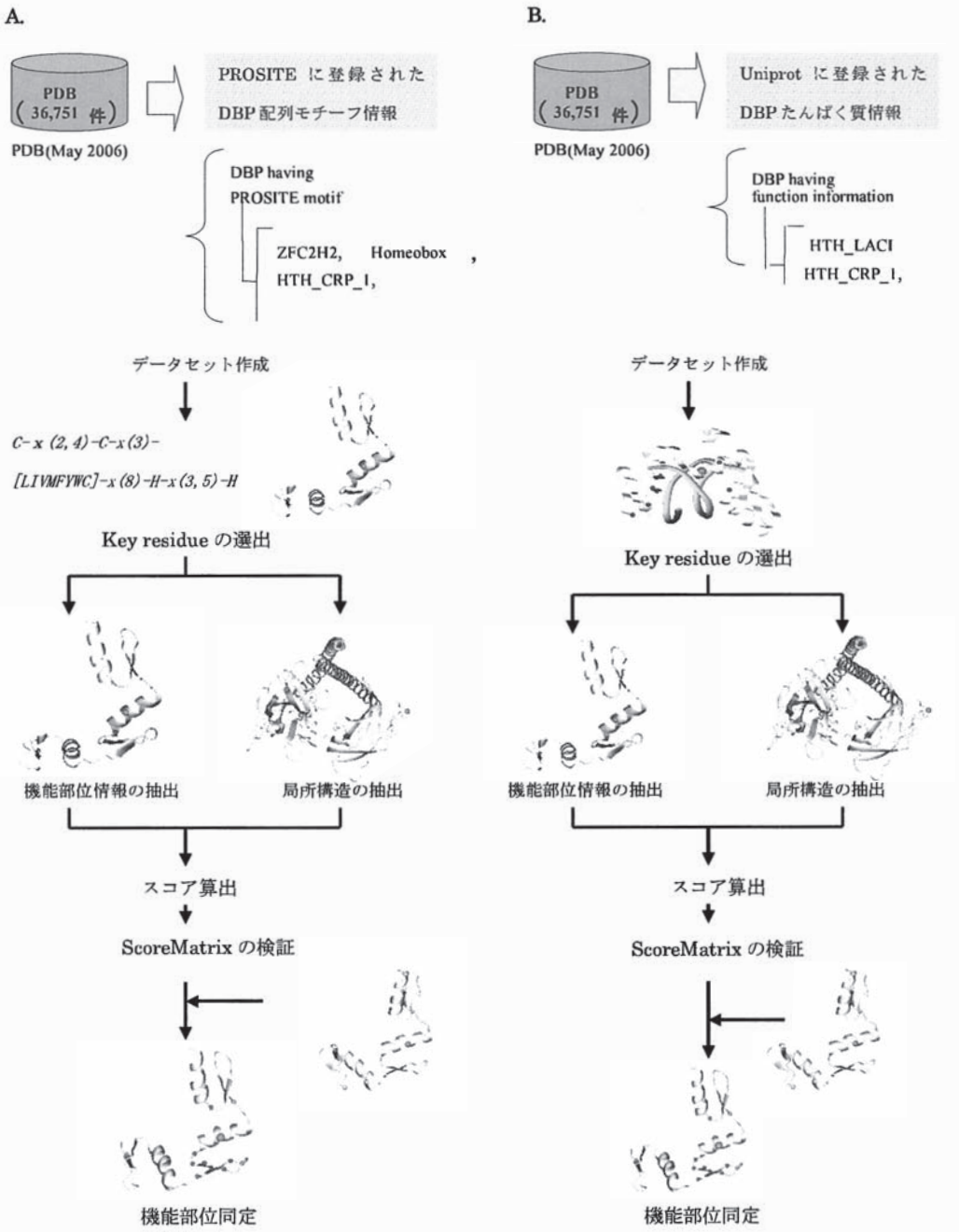

Figure6. Scheme of FCANAL method

A. 配列モチーフ情報による機能部位予測：PROSITE pattern の中から最も高度に保存されているアミノ酸残基を１つ選出し、それを key residue（ピンク球）とした。

B. DNA 結合部位の機能部位予測：DNA に結合しているアミノ酸残基(赤い球)から重心を算出し、これに最も近いアミノ酸残基を key residue（ピンク球）とした。

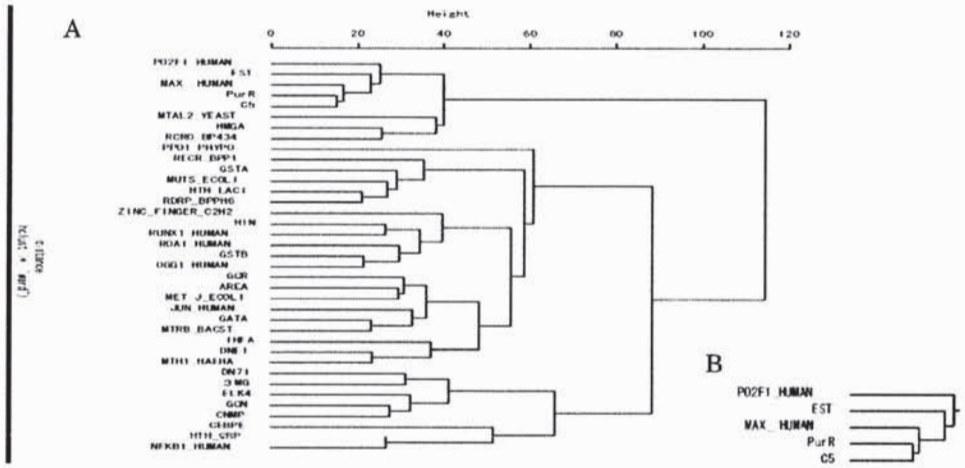

Figure7.　全 DBP のクラスタリング結果
(A)　全レコードを用いたクラスタリング結果
(B)　クラスタ 1

PO2F1_HUMAN の立体構造 5 個、EST の立体構造 6 個、MAX_HUMA N の立体構造 4 個、PurR の立体構造 21 個,C5 の立体構造 16 個をまとめて 1 つのデータセットとした。更に、Key Residue は、前項と同様の手法で選出し、Ser をそれと定め、FCANAL を実行した。その結果、F 値 2.00、A 値 2.82 を示し、機能部位予測精度は非常に良好だった。また、クラスタ 1 の Scores Data においても、DNA 結合部位の局所構造とその他の局所構造が二極化された(Figure8)。つまり、この Scores Data から、ただ DB 関数の傾向が似ていただけではなく、機能部位と立体構造 Domain は無関係という我々の仮説が正しいことが証明された。また、Score Matrix より得られた機能部位情報を基に、各 DBP の立体構造に機能部位をマッピングした(Figure9)。Key Residue から

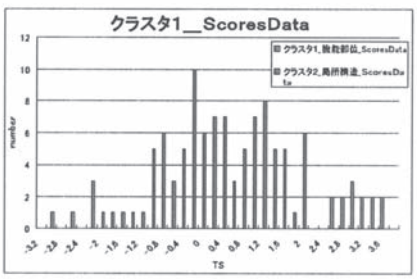

Figure8　クラスタ 1 の Scores Data

一定の距離には、共通のアミノ酸残基が高度に保存されていた。これより、複数の DBP の DNA 結合部位において共通の空間配置に存在するアミノ酸残基が、転写調節に何らかの影響を与えている可能性が考えられる。

6.　おわりに

本研究では、DBP 機能部位のアミノ酸残基の空間的出現確率に基づいて数量化し、DBP の DNA 結合部位の特徴づけを行った。DBP の DNA 結合部位は、高度に保存されているため、空間位置にしめる各アミノ酸残基の存在確率は相対的に高いものであると考えられる。そこで、我々は DNA 結合部位の局所構造情報に基づいた機能部位予測を目的に開発された FCANAL を応用し、DBP の機能部位予測を行った。当初、機能部位の指標として PROSITE に登録されている配列モチーフ情報を引用した。しかし、立体構造形成に関与する配列モチーフと DNA 結合に関与する配列モチーフが混在していた為、FCANAL による機能部位予測精度は非常に悪いものが含まれていた。そこで、我々は DNA から 3.1Å以内に存在するアミノ酸残基を抽出し、それらの重心位置を活性化中心とみなして、FCANAL に DNA 結合部位を再学習させてみた。その結果、DNA

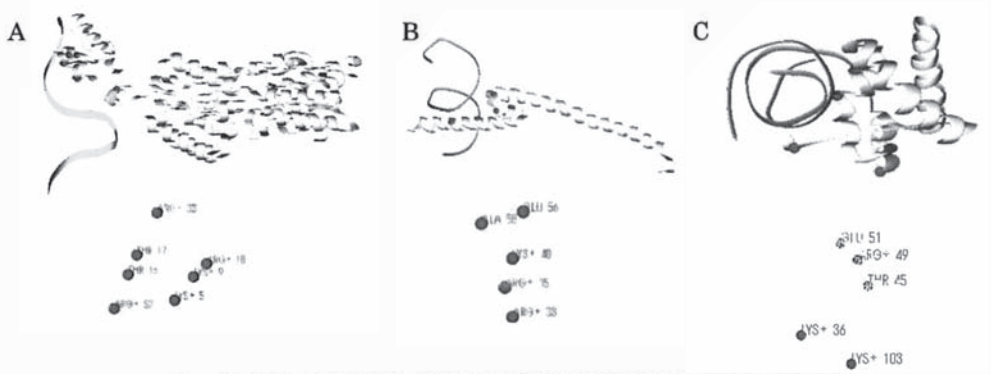

Figure9　HTH_LACI(A),MAX_HUMAN(B),PO2F1_HUMAN(C)の立体構造と機能部位

(A)　HTH_LACI は Helix-turn-Helix の Domain を持つ」(PDBID：1bdh)
(B)　MAX_HUMAN は、Leucin_zipper の Domain を持つ（PDBID:1an2）
(C)　PO2F1_HUMAN は、Homeo_Box の Domain を持つ(PDBID:1oct)
＊いずれもピンク球は FCANAL で予測した機能部位を構成するアミノ酸を示す」

結合部位におけるアミノ酸残基配置の空間的出現確率が、構造特徴によく対応することを見出した。これより、機能と機能部位の対応づけが完了した。次に機能部位情報を用いてクラスタリングを行い、DBP の機能による分類を試みた。その結果、DBP の立体構造形成に大きく関与している立体構造 Domain と実際に機能を発現する部位の間に相関関係は見られなかった。一方、異なる立体構造 Domain 間でもって、DBP の DNA 結合に関与すると考えられるアミノ酸残基は保存されていた。この結果は、DBP とシスエレメント配列との結合に何かしらの法則がある可能性を示唆している。

謝辞

本研究を遂行するにあたり、熱心にご指導くださいました、東京理科大学薬学部生命情報科学研究室の鈴木智典助教に深く感謝致します。また、研究の相談に乗って下さいました、東京理科大学大学院薬学研究科生命情報科学研究室の西浜睦子、石田秀徳、伊藤麻希子に心より感謝致します。

文献

1)　L.Narlikar　and　Alexander.A.hartemink, oxfordjournal-BIOINFORMATICS,2,157-163 (2006)

2)T.Asaoka ,T.Ando ,T.Meguro and I.Yamato, CBIJ, 3,96-113(2003).

3)A.Suzuki, T.Ando, I.Yamato and S.Miyazaki, CBIJ, 3,39-55(2005).

4) H.M.Berman, J.Westbrook, Z.Feng, G.Gilliland, T.N.Bhat, H.Weissig, I.N.Shindyalov, P.E.Bourne The Protein Data BankNucleic Acids Research, 28, 235-242 (2000) PDB

5) A Bairoch, P Bucher and K Hofmann, Nucleic Acids Research,1, 217-221(1997) PROSITE

6)半田宏,和田忠志,山口雄輝,転写研究　集中マスター,73-81 羊土社(2005)

7) A.Bairoch, R. Apweiler, Cathy H. Wu, Nucleic Acids Research, 33,(2005) UNIPROT

8)松原謙一,ゲノム情報生物学,100-113 中山書店(2003)

9)村上康文,古谷利夫,バイオインフォマティックスの実際,71-111,講談サイエンティフィク

10)西田英郎,実例クラスター分析,101-147,内田老鶴圃,(1992)