転写因子の比較ゲノム情報解析：原核型との対比における真核型転写因子の差異と特性

●西川 建[1)]　◆福地 佐斗志[2)]
1) 前橋工科大学　2) 国立遺伝学研究所

＜研究の目的と進め方＞

我々はこれまでに、ゲノム情報の中から転写因子のみを自動的に判別し収集する方法を開発してきた。その結果、転写因子に関する種横断的な比較ゲノム解析が可能となった。これまでに、真正細菌に比べて古細菌の転写因子ファミリーは少なく、前者が後者を含む形の「ほぼ包含関係」にあること、真核生物の転写因子の多くのものは長大な不規則（intrinsic disorder）領域を含むことを明らかにした。すなわち、原核生物においては古細菌を含めて転写因子の共通性、進化的連続性が認められるが、原核生物と真核生物の転写因子の間では、サイズの違い、DNA 結合ドメインの違い、不規則領域の有無など、量的・質的な違いが存在し、両者の異質性・進化的不連続性が認められた。

本研究では、ヒト転写因子に的を絞り、真核生物に特徴的な不規則領域について、以下のような観点から、その同定法の見直しを行う。一般に、不規則領域を持つタンパク質の分子構成は、構造ドメインと不規則領域に２分されるはずである。ところが、これまでの予測法では既知のドメイン、不規則領域、空白（予測できない）領域へと３分割された。空白領域は全長の 20％程度に達するが、本来この部分は不規則領域と未知の構造ドメインからなると考えられる。したがって、未知ドメインと不規則領域を判別する方法が開発できれば、タンパク分子の全体を不規則領域とドメイン（既知と未知を含む）に２分できることになる。それと同時に、構造未知ドメインの位置も予測されるので、転写因子における未知ドメインの割合を知ることができる。

＜ 2007 年度の研究の当初計画＞

昨年度までのヒト転写因子の解析において、ドメイン／不規則領域のどちらにも判別されずに残された空白領域（20％）の判別を行うことを今年度の目標とした。一般に、通常の不規則領域予測は不規則領域に特徴的なアミノ酸組成の偏りに着目して予測を行うが、構造ドメインと不規則領域を比較した場合、もう１つの顕著な違いとして配列保存性の差異がある。すなわち、ドメインの配列保存性は高く、原核／真核生物の分類枠を越えて微弱なホモロジーが検出できることが知られている。一方、不規則領域の配列保存性は極端に悪く、ヒトを基準にすると、哺乳類の範囲では配列アラインメントが可能であるが、魚類くらいまで離れるとアラインメントが不確かになり、それ以上遠縁の関係になるとホモロジーがまったく検出されないことが多い。不規則領域の変異性が高くなる理由は、球状構造を形成しないため構造的な制約による淘汰を受けないからだと理解できる。このような保存性（変異性）の大きな差異を利用すれば、ドメインと不規則領域が区別できると期待される。

昨年度までと同じく、398 個のヒト転写因子を解析の対象とする。個々の転写因子の構造ドメインの位置や、予測プログラムによる不規則領域の位置に関しては GTOP データベース（国立遺伝学研究所より公開）を参照する。GTOP では、ゲノム既知の生物種のもつ全タンパク質を対象に、プロファイル型ホモロジー検索ツール（PSI-BLAST, HMM）による構造ドメイン（SCOP, PDB）と機能ドメイン（Pfam）の同定を行い、その結果を情報提供している。また、不規則領域に関しても、予測プログラムとして定評のある DISOPRED2 を用いた予測結果が提供されており参照できる。

＜ 2007 年度の成果＞

不規則性タンパク質を集めた DISPROT データベースを参照して、ドメイン／不規則領域に対する判別分析を行うための学習セットとテストセットを用意した。判別分析には、BLAST 検索による配列保存性、アミノ酸組成、二次構造傾向性（PSIPRED のスコア）、の３つの要因を考慮して判別プログラムを作成した。学習セットを用いてパラメータの最適化を行い、テストセットを用いて判別能を調べたところ、89％の正答率を得た。ただし、転写因子に適用するときには、GTOP を参照して得られる構造ドメインと不規則領域を優先的にアサインし、残りの空白領域に対してのみ判別分析を用いた。

この方法をヒト転写因子に適用した結果を図１（左）に示す。

全体的な平均値として、構造ドメイン 40％（既知ドメイン 35％、未知ドメイン 5％）、不規則領域 58％、どちらとも判別されない空白領域 2％となった。この空白領域は十分量のホモログが得られない場合に生じた。昨年度までの結果と比べると、以前の空白領域のうち、56％が不規則領域、32％が未知ドメインと判別され、残りの 12％は空白のままであった。また、比較のために大腸菌の転写因子（136 個）にも同様の方法を適用した結果を図１（右）に示す。大腸菌にも昨年度までは 20％程度の空白領域が残されていたが、今回そのほとんどは未知ドメインと判別され、不規則領域は全体の 3% 程度しかないことが示唆された。図１を見て明らかなように、同じ転写因子といっても、原核生物と真核生物ではその分子構成に大きな違いがあることが再確認された。

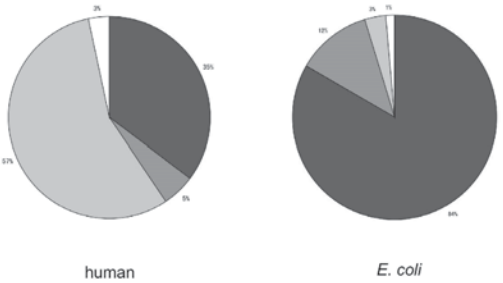

図１. ヒト（左）および大腸菌（右）転写因子の構造ドメイン（赤、構造既知；橙、構造未知）と不規則領域（灰色）の比較

このようにして、ヒト転写因子では全長の６割近い領域が不規

則領域であると判明するとともに、いくつかの構造未知ドメインが同定された。そのうちの１例を図２に示す。このタンパク質は、SREBP-1a (Sterol regulatory element-binding protein 1a) と呼ばれ、脂質代謝、特にコレステロール、脂肪酸などの代謝に関連する遺伝子群を転写調節する転写因子である。SREBP ファミリーは、この他に SREBP-1c, SREBP-2 が知られており、ともに同様のドメイン構成をしている。図２において、１番上がスケール、２番目が昨年度までのアノテーション（緑、構造ドメイン；グレー、不規則領域）、３番目は本年度のアノテーション（赤、構造ドメイン；橙、構造未知ドメイン；グレー、不規則領域；青、膜貫通部位）、４段目は論文等に見られるドメイン構成である。昨年度までの判別では、C 末端側の大部分は短い不規則領域が少し予測された以外は空白領域で占められていたが、今回の結果では、この部分に構造未知ドメインが予測された。論文等を参照してみると、この C 末端領域は carboxyl regulatory domain と呼ばれ、SREBP cleavage activating protein (SCAP) の WD ドメインと複合体を形成することが知られている。橙の領域は４つ見られるが、これは必ずしも４つの構造ドメインが存在することを示してはいない。我々の開発したプログラムは、ドメイン／不規則領域を判別するのみで、構造単位の切れ目を言い当てるものではない。その意味で、この regulatory ドメインに見られる不規則領域はドメインどうしをつなぐリンカーかもしれない。いずれにせよ、C 末端側は全体として構造未知のドメインを形成している可能性が高そうである。

ヒト転写因子には、Trans-activation domain (TAD) とよばれる機能部位が実験的に確認されている場合が多い。これらの領域は、転写因子が DNA に結合した後、転写装置を活性化するのに重要と考えられており、多くの場合、不規則領域中に存在する。これらの TAD 領域は、単独ではフォールドしないが、相互作用相手と結合し構造を作ることが知られている。SREBP にも N 末端領域に TAD の存在が知られている（図中、TA）。この領域は、昨年までの結果では空白だったが、今回は不規則領域と判別された。このことは、この TAD も上記のような様式での転写活性化を行っている可能性を示唆するものである。

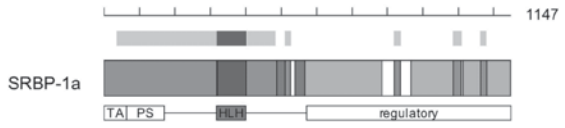

図２ ヒト転写因子 SREBP-1a の分子構成

＜国内外での成果の位置づけ＞

タンパク質の不規則領域の情報解析は、K. Dunker（Indiana 大学）らグループによって、2000 年前後から精力的に行われてきた。ヒト転写因子に関しても、彼らは我々と同時期に解析結果を発表しているが、Dunker らの解析は不規則領域を予測するだけで、ドメインとの関係を一切考慮しないため、小さい DBD と長大な不規則領域から成るという転写因子の特徴を捉えきれていない。一方、本研究の結果は、ヒト転写因子のほぼ全域に関してドメイン／不規則領域の判別を行ったものであり、全体の統計や、個々のタンパク質の分子構成に関する成果は他に類を見ないものといえる。現在、投稿論文を準備中であり、また個々の転写因子のドメイン構成に関してもインターネット上で公開する予定である。

＜達成できなかったこと、予想外の困難、その理由＞

今回開発した判別分析は、転写因子を query として BLAST 検索を行い、得られたホモログ配列を入力データとする方法である。そのため、最終的な判別能は得られるホモログ配列の多少に依存し、少しのホモログ配列しか得られない場合や、ホモログ配列がまったく存在しないときは判別不能となってしまう。幸いにも、転写因子を対象とする限りは比較的にホモログが多く見つかる場合がおおいので、図１の空白領域で示されるように、その割合は数パーセントに留まった。しかしながら、一般のタンパク質に適用するときには、種特異的なタンパク質なども含まれると想定されるので、判別不能の割合が増加する恐れがある。

＜今後の課題＞

本研究で開発した判別分析の方法は、転写因子ばかりではなく、タンパク質一般にももちろん適用可能である。従来の方法では、ヒトの全タンパク質中に 40%、大腸菌では 45% 程度の未知領域が残されており、今回の方法を適用すれば、これらの領域にどれほどの未知構造ドメインがあるか、不規則領域はどの程度の割合か、を知ることができそうである。本研究の成果を見れば、最後に残される空白領域は数％と予想されるため、原核生物では長大な不規則領域は皆無に近いだろうと言われてきた点を、全領域にわたる判別を用いることによって初めて確認することができるだろう。また、全域にわたる判別を用いれば不規則領域と Alternative Splicing Variant (ASV) の関係も、より明確なかたちで解析できるだろう。本年度は、ドメインと不規則領域の判別に注力したが、不規則領域には多くの TAD やその他の機能部位が埋もれている可能性が高い。これら機能性部位の発見も今後に残された重要な課題と考えられる。

＜成果公表リスト＞

1) 論文／プロシーディング

1. 0705021227
Minezaki, Y., Homma, K., and Nishikawa, K.: Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. J. Mol. Biol. 368(3), 902-913 (2007)

2) データベース／ソフトウェア

1. 0507070143
GTOP データベース
http://spock.genes.nig.ac.jp/~genome/gtop.html