バイオインフォマティクスによる配列データベース横断解析に基づく機能上流ORFの推定

髙橋　広夫[1]*・伊藤　素行[2]・尾之内　均[3]

ヒトゲノムの98%がタンパク質をコードしない非コード領域であり，非コード領域の多くがゲノムのジャンク領域と考えられていた．しかし，2012年9月8日号のNatureの表紙を飾ったEncyclopedia of DNA Elements（ENCODE）プロジェクトにより，ヒトゲノムの80%は機能を持った領域であることが明らかにされた[1]．近年，ゲノム非コード領域のさまざまな生命現象への関与が注目を浴びており，本稿ではmRNAの5′非翻訳領域（5′-UTR）に存在し，短い機能性ペプチドをコードする可能性がある上流ORF（upstream open reading frame：uORF）を取り上げて紹介する．

上流ORFと翻訳アレスト

真核生物では，mRNA上にタンパク質に翻訳されない非翻訳領域（untranslated regions：UTR）があり，生物学的意義は軽視されてきた．ところが近年，5′-UTRには，uORFと呼ばれる短いペプチドをコードする読み枠が多く存在することが注目を浴びている[2–5]．図1Aのように，リーキースキャニング（leaky scanning）やリイニシエーション（reinitiation）と呼ばれるメカニズムにより，uORFがあっても下流のmain ORF（mORF）が翻訳される場合も多いが，図1Bに示したようにuORFが翻訳されることによって下流のmORFの翻訳に影響を与える機構（releaseやstall）があり，特にuORFの翻訳の途中でリボソームを停滞（stall）させることで翻訳制御やmRNA分解制御に関わる制御性ペプチドが注目され始めた[6–9]．このような制御性uORFのいくつかは，翻訳途中の新生鎖がリボソーム出口トンネルの狭窄部位と相互作用することで，リボソームを停滞させると考えられている[10]．このようなリボソームの停滞は，後続のリボソームのスキャニングをブロックすることで，リボソームの渋滞を引き起こし，結果として下流のタンパク質コード領域であるmORFの翻訳を抑制する[11]．このようなリボソームの停滞は翻訳アレストとも呼ばれる．翻訳アレストがどのようなトリガーによって引き起こされるのかは，多くの場合は不明であるが，細胞内のエフェクター（代謝産物など）に応答するuORFがいくつか報告されている．たとえば，アルギニンに応答するアカパンカビの*arg-2*のuORF[11]，ポリアミンに応答する動植物の*AdoMetDC*のuORFである[12–13]．このようなuORFは，進化的に保存されていることが多く，進化的に保存されたアミノ酸配列を持つuORFはConserved Peptide uORF（CPuORF）と呼ばれる．

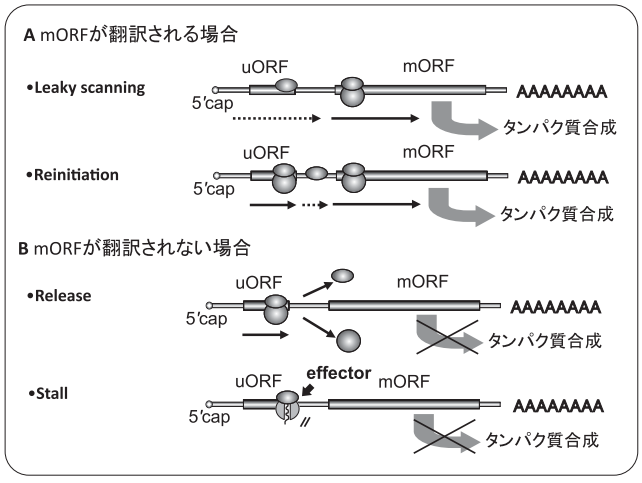

図1．uORFによるmORFの翻訳制御機構

情報工学に基づくCPuORFの同定

情報学的なCPuORFの網羅的同定は，植物に関しては，2007年に，シロイヌナズナとイネのオルソログの比較に基づいて50個（19ファミリー），シロイヌナズナのパラログを比較することによって14個（7ファミリー）同定された[14]．2008年には，4種のイネ科植物（小麦，大麦，トウモロコシ，モロコシ）の比較によって29個のCPuORFが同定された[15]．さらには，2012年に6種の双子葉植物（シロイヌナズナ，ブドウ，ダイズ，ワタ，オレンジ，タバコ）の比較によって，4個のCPuORFが同定された[16]．しかし，一般的にuORFの配列は短いため，このような少数の生物種間での比較では，解析に用いた生物種に存在するCPuORFであっても，それらの種間での配列の類似性が不十分なためにCPuORFとして同定されない場合がある．また，これまでの解析では，用いられる生物種は完全長cDNA配列情報が明らかになった種に限定されていた．そこで筆者らは，部分的なcDNA

著者紹介　[1]金沢大学 医薬保健研究域薬学系（准教授）　E-mail: takahasi@p.kanazawa-u.ac.jp
[2]千葉大学大学院薬学研究院　生化学研究室（教授），[3]北海道大学大学院農学研究院（准教授）

の配列情報であるexpressed sequence tag（EST）データベース（DB）をフル活用し，BLASTを用いて不特定種間での比較を可能にすることで，データベース横断的に解析できるBAIUCAS（BLAST-based algorithm for identification of uORFs with conserved amino acid sequences）法を開発した．BAIUCAS法をシロイヌナズナゲノムに応用することで，2012年にシロイヌナズナに比較的近縁な種でのみ保存されているものを含む18個（17ファミリー）のCPuORFを同定した[17]．BAIUCAS法で同定した18個のCPuORFのうち4個は，Vaughnらのグループがわずかに先に発表してしまったため，新規のものは14個であった．このように，植物では網羅的なCPuORFの同定が盛んに行われている．

一方で，動物では2006年にヒトとマウスとのcDNAの比較解析から，204個のヒトCPuORFと198個のマウスCPuORFが同定された[18]．また，2008年にショウジョウバエとその他のハエ目の比較から44個のCPuORFが同定された[19]．しかし，これ以降，CPuORFの網羅的同定の報告がなく，ヒトゲノム情報の日々のアップデートにより，UTR情報が大幅にアップデートされたため，Croweらの同定した204個のうち39個は，最新の5′-UTR情報から消えてしまった．また，執筆時点で，Croweらの同定したCPuORFの情報が記載されたサプリメントデータは，ネット上から削除されてしまい，参照不可能である．このように，ヒトをはじめとした動物CPuORFの情報は限定的である．

アミノ酸配列依存的な機能CPuORF

アミノ酸配列依存的にmORFの翻訳を抑制するCPuORFとして，前述のアカパンカビの*arg-2*，動植物の*AdoMetDC*の他，植物の*bZIP11*[20]，ヒトの*CHOP*[21]と*RARβ2*[22]のCPuORFが報告されていたが，少数例しか報告されていなかった．そこで，筆者らは，BAIUCAS法で同定したシロイヌナズナの17ファミリーに着目し，このうち，16ファミリーのCPuORFに関してアミノ酸配列依存的な翻訳抑制機能の有無を一過性発現系で検証した[23]（図2）．筆者らの解析対象外とした*AtTTM3*のCPuORFにコードされるペプチドは，酵母や動物のCDC26と相同であり，リボソームからリリースされてから機能すると予測され，最近になって予想通り植物の細胞周期の制御に関わることが報告された[24]．筆者らは，図2Aの示すように，構成的プロモーター（カリフラワーモザイクウイルス35S RNAプロモーター：P35S）とウミシイタケルシフェラーゼ（Rluc）の間に野生型5′-UTRを入れた野生型コンストラクト（WT）と，CPuORFのサイズは変えずにフレームシフト変異でアミノ酸配列を変えたコンストラクト（fs）を作製し，シロイヌナズナの培養細胞株MM2dのプロトプラストにエレクトロポレーションで導入して，一過性発現系でレポーターの比活性を調べた（図2B）．その結果，11個のCPuORFがアミノ酸配列依存的な翻訳抑制機能を持っていることが示唆され，このうち2倍以上の比活性を示した*ANAC082*，*CIPK6*，*OTLD1*，*AT3G15430*，*AT5G27920*のCPuORFペプチドについては，アラニンスキャニングによって翻訳抑制に関わるアミノ酸残基を同定した．シロイヌナズナ*ANAC096*のCPuORFを用いた実験ではWTとfsのレポーター活性の差はわずかであったが，トマトの*ANAC096*ホモログのCPuORFはアミノ酸配列依存的な翻訳抑制機能を示し，アラニンスキャニングによって翻訳抑制領域を決定した[25]．つまり，BAIUCAS法で同定した16個中の少なくとも6個は，アミノ酸配列依存的な翻訳抑制機能を持っていることが示された．

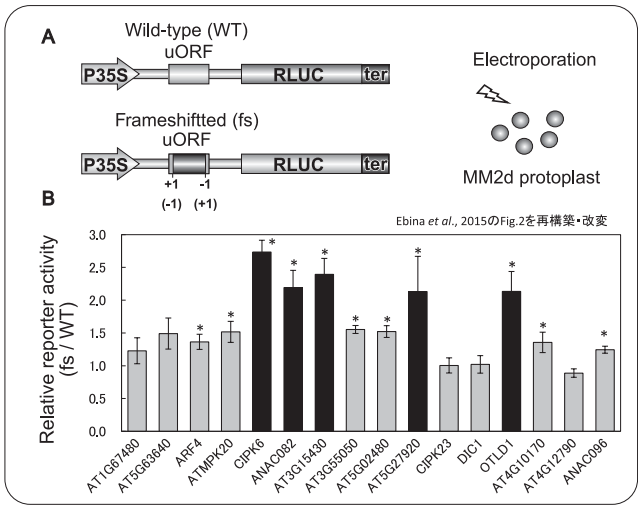

図2．一過性発現系を用いた翻訳制御機能の検証

BAIUCAS法の限界と拡張

筆者らが開発したBAIUCAS法は，シロイヌナズナから新規CPuORFを同定するのに有用であったが，いくつかの問題点があった．①シロイヌナズナのゲノムDBであるThe Arabidopsis Information Resource（TAIR）[26]のデータ形式に合わせて構築したアルゴリズムであったため，他の生物へ応用が容易ではなかった．②ESTのみを対象としたアルゴリズムであったが，次世代シーケンサーの発展により，近年，急速にデータが蓄積しつつあったtranscriptome shotgun assembly（TSA）には，対応していなかった．③種分類DBであるTaxonomy DBには，

対応していなかったため，種を細かく分類したカテゴリーを定義するのが手作業となり，BAIUCAS法の他の生物への応用の障害となっていた．④ゲノムDB上では，ある生物ではCPuORFのように見えるものが，他の生物ではタンパク質の一部になっているような『偽CPuORF』を除外する機能がなく，BAIUCAS法の解析後，偽CPuORFを手動で除く必要があった．⑤CPuORFペプチドのアミノ酸配列が保存されているかどうかを判定するK_a/K_s値の算出やK_a/K_sシミュレーションには，手動で選んだ配列を用いていた．⑥BAIUCAS法におけるBLAST解析は，国立遺伝学研究所のDNA Data Bank of Japan（DDBJ）[27]のsimple object access protocol（SOAP）というwebシステムを使っていたが，DDBJのサーバーシステムの入れ替えの際に，このサービスが停止され，BAIUCAS法によりさらなる解析が不可能になってしまった．このように当時としては，最大の検出力を誇るBAIUCASであったが，さまざまな問題と課題があった．

そこで，BAIUCAS法を改良・拡張することにした．①については，真核生物のゲノム情報を同一形式で提供しているENSEMBL[28]に対応することで，ENSEMBLに登録されている生物のゲノムを，自由に解析できるようになった．②については，TSAのデータも扱えるようにし，また，扱うデータが多くなったため，重複する配列はアセンブルすることで冗長性を下げ，DBサイズを縮小し，解析しやすい独自の配列DBを構築した．③については，配列DBから，種分類DB（NCBI Taxonomy DB）へシームレスに解析できるように，配列とリンクした種分類DBを構築した．この結果，生物界を細かく定義付けすることが可能となり，CPuORFの進化保存範囲を詳細に決定することが可能となった．④については，融合率（uORFが他の生物のmORFの一部になっている可能性）を計算するアルゴリズムを開発した．⑤については，K_a/K_s値の算出やK_a/K_sシミュレーションで使うuORF配列を自動決定するアルゴリズムを開発した．⑥については，ローカルコンピュータでBLAST解析できるようにアルゴリズムを改良し，クエリレベルの並列化を採用した結果，大幅な高速化を実現した．このようにプログラム的にいろいろな最適化も行い，最終的に図3のような長大なパイプラインに基づくEvolutionary search for upstream open reading frames with conserved amino acid sequences（ESUCA）法が完成した[29]．

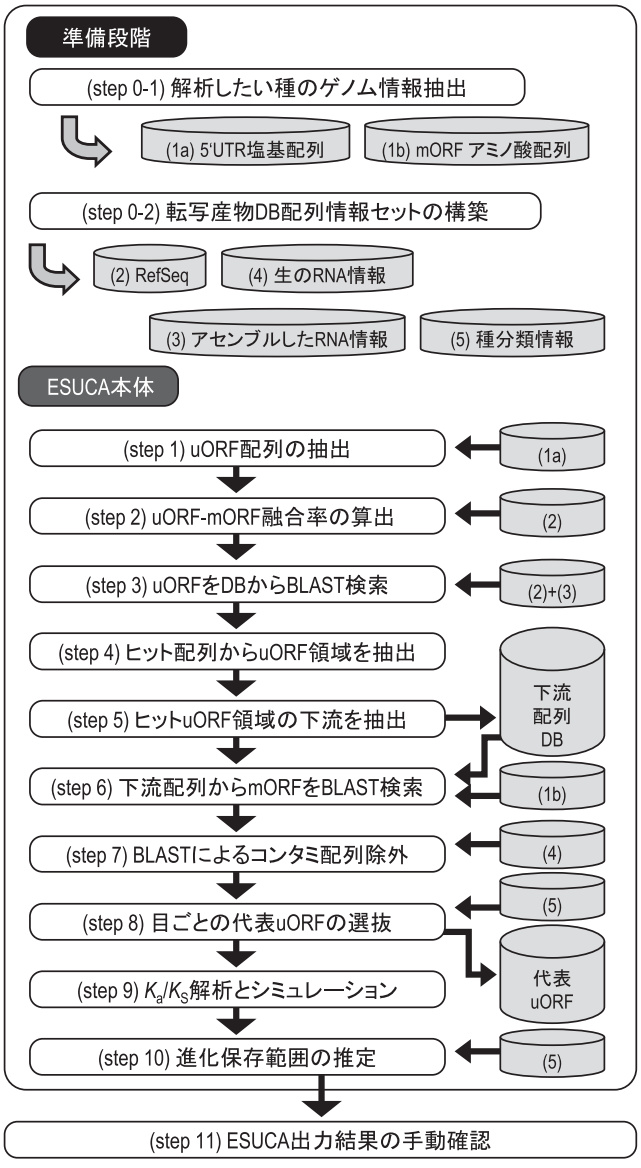

図3．ESUCA法アルゴリズム概要

ESUCA法を用いた動植物界からのCPuORFの網羅同定

ESUCA法は，容易にさまざまな生物のゲノムに応用できることから，植物では，5種の植物ゲノム（シロイヌナズナ，トマト，ポプラ，ブドウ）に応用したところ，150個近くのCPuORFファミリーを同定した．このうち，約半数が新しいファミリーである．このことは，ESUCA法で近縁種の比較ができるようになったことや，シロイヌナズナゲノムには保存されていない，ポプラやブドウゲノムを解析することで初めて見つかったCPuORFのファミリーが大量に同定できたためである．一方，動物ゲノムについては，4種（ショウジョウバエ，ゼブラフィッシュ，ニワトリ，ヒト）のゲノムについて解析したところ，合わせて1000以上のCPuORFファミリーが同定された．動植物それぞれのCPuORFについて一過性発現系を用いて翻訳抑制機能を検証したところ，比較的近縁な種間でしか保存されていないCPuORF

でもアミノ酸配列依存的な翻訳抑制機能を持ちうることが証明できた（未発表データ）．

おわりに

これまでの知見や筆者らの解析から，多くのCPuORF型の翻訳アレスト配列が見つかってきた．また，本稿では，AUGから翻訳されるCPuORFについてのみを紹介したが，翻訳はAUG以外からも始まることが，近年分かってきている．筆者らはESUCAを拡張することで，予備的ながら，非AUG型のCPuORFの同定にも成功している．今後，このような配列をさらに解析することで，アレスト配列の規則性やアレスト機構の解明が期待される．このような発見は，タンパク質領域における翻訳アレスト領域をインフォマティクス的な方法から予測することを可能とするかもしれない．

文　献

1) Dunham, I. *et al.*: *Nature*, **489**, 57 (2012).
2) Churbanov, A. *et al.*: *Nucleic Acids Res.*, **33**, 5512 (2005).
3) Galagan, J. E. *et al.*: *Nature*, **438**, 1105 (2005).
4) Kawaguchi, R. and Bailey-Serres, J.: *Nucleic Acids Res.*, **33**, 955 (2005).
5) Rogozin, I. B. *et al.*: *Bioinformatics*, **17**, 890 (2001).
6) Cruz-Vera, L. R. *et al.*: *Curr. Opin. Microbiol.*, **14**, 160 (2011).
7) Ito, K. and Chiba, S.: *Annu. Rev. Biochem.*, **82**, 171 (2013).
8) Morris, D. R. and Geballe, A. P.: *Mol. Cell Biol.*, **20**, 8635 (2000).
9) Somers, J. *et al.*: *Int. J. Biochem. Cell Biol.*, **45**, 1690 (2013).
10) Bhushan, S. *et al.*: *Mol. Cell*, **40**, 138 (2010).
11) Wang, Z. and Sachs, M. S.: *Mol. Cell Biol.*, **17**, 4904 (1997).
12) Law, G. L. *et al.*: *J. Biol. Chem.*, **276**, 38036 (2001).
13) Hanfrey, C.: *J. Biol. Chem.*, **280**, 39229 (2005).
14) Hayden, C. A. and Jorgensen, R. A.: *BMC Biol.*, **5**, 32 (2007).
15) Tran, M. K. *et al.*: *BMC Genomics*, **9**, 361 (2008)
16) Vaughn, J. N. *et al.*: *RNA*, **18**, 368 (2012).
17) Takahashi, H. *et al.*: *Bioinformatics*, **28**, 2231 (2012).
18) Crowe, M. L. *et al.*: *BMC Genomics*, **7**, 16 (2006).
19) Hayden, C. A. and Bosco, G.: *BMC Genomics*, **9**, 61 (2008)
20) Rahmani, F. *et al.*: *Plant Physiol.*, **150**, 1356 (2009).
21) Jousse, C. *et al.*: *Nucleic Acids Res.*, **29**, 4341 (2001)
22) Reynolds, K. *et al.*: *J. Cell Biol.*, **134**, 827 (1996)
23) Ebina, I. *et al.*: *Nucleic Acids Res.*, **43**, 1562 (2015)
24) Lorenzo-Orts, L. *et al.*: *Nat. Plants*, **5**, 184 (2019).
25) Noh, A. L. *et al.*: *Plant Biotechnol.*, **32**, 157 (2015).
26) The Arabidopsis Information Resource: https://www.arabidopsis.org/ (2019/5/30).
27) 生命情報・DDBJセンター：https://www.ddbj.nig.ac.jp/ (2019/5/30).
28) Ensembl: https://www.ensembl.org/ (2019/5/30).
29) Takahashi, H. *et al.*: *bioRxiv*. 524090 (2019).