論文の内容の要旨

論文題目

Finding of novel short coding sequences from human full-length cDNAs by mass spectrometry

和訳

質量分析計によるヒト完全長 cDNA 配列からの新規小翻訳領域の発見

指導教官　菅野　純夫　助教授

東京大学大学院医学系研究科

平成 12 年 4 月進学

医学博士課程

病因・病理学専攻

氏名　尾山　大明

（序論）

大腸菌や酵母の全ゲノム配列の解読完了に続き、ヒトゲノムに関しても全配列の解読完了が宣言された。ヒトに関しては、ゲノム配列の解読と並行して完全長 cDNA に関する情報の蓄積も著しく進展し、それに基づいたタンパク質コード領域に関する情報の蓄積及び整理も進行している。現在、完全長 cDNA 情報に基づいて整理された代表的なタンパク質データベースである NCBI の Reference Sequence (RefSeq)タンパク質データベース(2003 年 3 月現在)によると、当データベースに登録されている全タンパク質 (17,280 個)の 96.2 % (16,609 個)は、ORF の長さが 100 アミノ酸残基よりも長いものであり、100 アミノ酸残基以下の ORF がコードする低分子タンパク質の数は非常に少ないと考えられている。

しかしながら完全長 cDNA に関する大規模な配列解析は、非常に多くの未知低分子タンパク質が存在する可能性を示唆している。まず 1 つ目として、我々の完全長 cDNA プロジェクト(FLJ プロジェクト)において、RefSeq に登録されている cDNA に該当しない新規の完全長 cDNA 配列が 1 万種以上得られているが、興味深い事にその半数以上の cDNA 配列は 100 アミノ酸残基よりも長い ORF を持たない。これらの完全長 cDNA が短いタンパク質をコードしている可能性が十分考えられる。

更に 2 つ目として、代表的な約 5,000 の完全長 cDNA に関して 5' 端の非翻訳領域に関する解析を行ったところ、これらのほぼ半数が開始コドンの上流に少なくとも 1 つの ATG コドンを持つことが報告されている。真核生物の典型的な翻訳のメカニズムにおいては、リボソーム前駆体が mRNA の 5' 端から 3' 端に向かってスキャンをすることによって開始コドンを探すことから、上記の配列解析の結果は既存のタンパク質コード領域の上流に多数の潜在的な低分子タンパク質コード領域が存在することを示唆している。

細胞の生命活動を理解する為には、細胞中で実際に発現し、機能を担っているタンパク

質群の全体像を把握する事が必要不可欠である。そこで当研究においては、高感度のnanoflow LC-ESI-MS/MS systemによる検出を基盤として低分子タンパク質を広範囲で同定する実験系を確立し、ヒトK562細胞をモデルとしてこの細胞中で実際に発現している低分子タンパク質を対象とした網羅的な解析を行ってみた。得られたMS/MSスペクトルからタンパク質を同定する際に、RefSeqのタンパク質データベースに加え、FLJ及びRefSeqの膨大な完全長cDNAデータに対して検索を行うことによって、既知及び未知の双方の低分子タンパク質に関する発現情報の整理を行った。その研究結果に関して以下に報告する。

(実験方法)

培養したヒトK562細胞中で発現している低分子タンパク質をLC-ESI-MS/MSを用いてより網羅的に探索する為に、細胞のlysateから低分子タンパク質を濃縮した試料を作製する事が必要となる。当研究においては2通りの方法で低分子タンパク質の濃縮画分の調製を行い、測定試料とした。

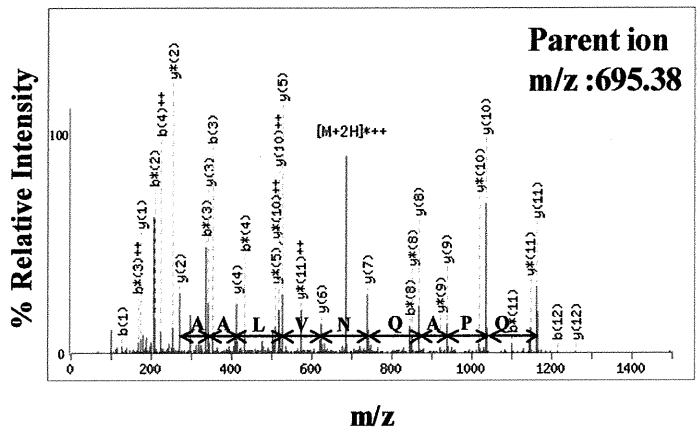
1つ目の方法として、まずヒトK562細胞のlysateをSDS-PAGEによって分子量に応じて分離展開を行った。泳動レーン上で約17 kDa以下の低分子量に相当する部位のみを切り出し、このゲル内に閉じ込められている低分子タンパク質を解析対象とした。ゲル内でトリプシンの添加によりタンパク質をペプチドに断片化し、アセトニトリル溶液によってゲル内からペプチドを抽出した。減圧遠心によってアセトニトリルを除去した後にZipTip™ (C_{18})によってペプチドを選択的に回収、濃縮し、質量分析用のサンプルとした。

2つ目の方法としては、まず回収したヒトK562細胞を酸存在下でホモジナイズし、低分子タンパク質が濃縮された上清を取得した。分離した上清液からゲルろ過によって塩等のたんぱく質以外の低分子物質を除去した後に、得られたタンパク質濃縮画分をトリプシンの添加により溶液中で直接ペプチドに断片化した。酵素消化により得られたペプチドをZipTip™ (C_{18})によって同様に回収、濃縮し、質量分析用のサンプルとした。

上記の2種類の方法で作製したサンプルを高感度のnanoflow LC-MS/MS systemに導入し、サンプル中に存在するペプチドに関するMS/MSスペクトルを順次取得した。得られたMS/MSスペクトルに関してはMascotアルゴリズムに基づいてデータベース検索を行うことにより、各MS/MSスペクトルが由来するタンパク質の同定を行った。まずRefSeqのタンパク質データベースに対して検索を行うことによって、ヒトK562細胞中で発現している主なタンパク質に関する情報を収集・整理した。そして同データベースに対する検索において同定されなかったMS/MSスペクトルに関して、更にヒト完全長cDNAデータセットに対して検索をかけることによって、新規の低分子タンパク質コード領域の同定を試みた。cDNAデータセットとしては、FLJ及びRefSeq双方の完全長cDNAコレクションを準備し、前者からはFLJコレクションに特徴的な完全長cDNAがコードする新規の低分子タンパク質を、そして後者からはRefSeqに既に登録されているタンパク質コード領域以外の未知翻訳領域の同定を試みた。

(結果及び考察)

まず、RefSeqのタンパク質データベースに対して検索を行った結果、当データベースに登録されているORFの長さが100アミノ酸残基以下の671個のタンパク質(2003年3月現在)の中で、リボソームタンパク質やトランスポーター等の計52個のタンパク質が同定された。更に、ヒト完全長cDNAデータセットに対する検索からは、FLJデータセットから4つ、RefSeqデータセットからは3つ、計7つの新規コード領域(ORF：100アミノ酸残基以下)が同定された。



NM_015532 novel short CDS (86 a.a.)

MATPARAPESPPSADPALVAGPAEEAECPPPRQPQPAQNVLAAPR
LRAPSSRGLGAAEFGGAAGNVEAPGETFAQRKIHLQIARQR

図1

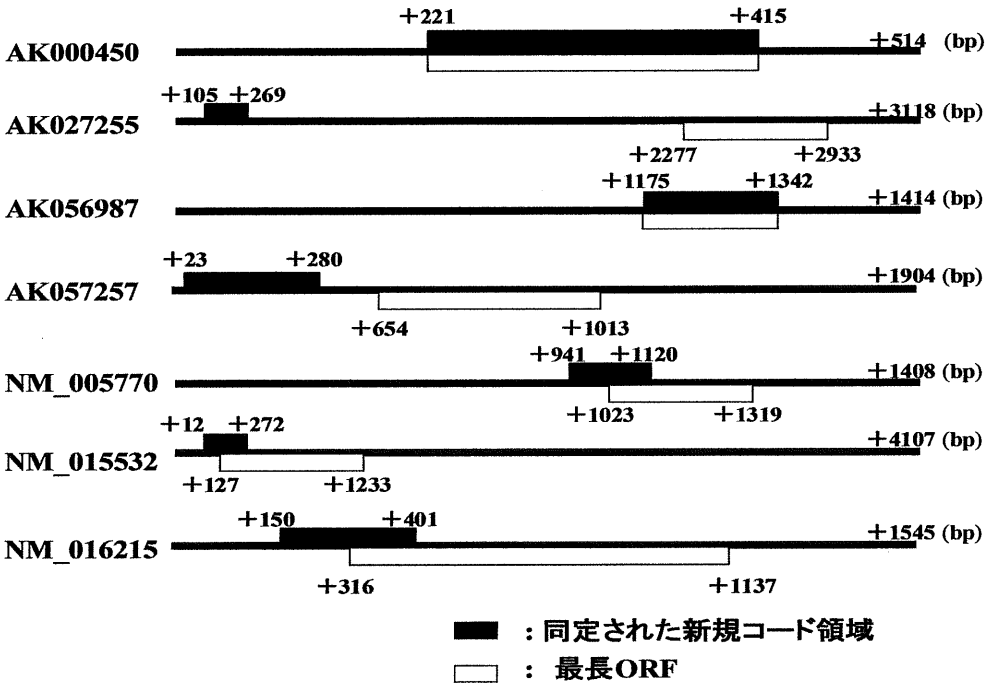


図2

図1に、新規低分子タンパク質由来のペプチドに対応するMS/MSスペクトルの1例を示す。興味深い事に、今回得られた7つの新規コード領域の中で、5つのコード領域は各完全長cDNA配列中で最も長いORFの上流に位置している事が分かった(図2)。下流の長いORFはタンパク質コード領域として既知あるいは強く推定されるものであり、これらの遺伝子は2つのコード領域を持ちうる事が示された。

また、最長ORFの上流に位置している5つのコード領域の開始コドンは全て、各完全長cDNA配列中で最も上流に存在していた。この結果は、リボソーム前駆体がmRNAの5'端から3'端に向かってスキャンをし、最初に遭遇した開始コドンから翻訳を開始するという典型的な翻訳開始のメカニズムの普遍性を支持するものである。RefSeqタンパク質データベースから同定された52個のタンパク質に関して、対応するRefSeqの完全長cDNA配列の情報に基づいて同様に開始コドンの位置を調べたところ、44個(85 %)のタンパク質が最も上流に位置するATGコドンから翻訳を開始している事が分かった。この解析結果は、タンパク質の翻訳が主にこのメカニズムに基づいて行われていることを改めて裏付けている。FLJ、RefSeq双方のcDNAデータセットの中には、配列中の最も上流に短いORFを持つcDNA配列が非常に多く存在することから、本研究の結果はこれらのORFがコードする多くの未知タンパク質が、実際に細胞中で翻訳されている事を示唆している。

当研究では、質量分析計による検出を基盤とした低分子タンパク質の探索を試みたわけだが、測定サンプルの調整法やLC systemの改良を行うことによって、より発現量の少ない低分子タンパク質まで検出対象を広げる事が出来ると考えられる。また、今回はヒトK562細胞を対象にして解析を行ったが、他の培養細胞や組織から調整したサンプルに関して測定を行うことにより、組織特異的な発現を示すタンパク質の探索を行うことも可能であると考えられる。

(まとめ)

ヒトK562細胞中で発現している低分子タンパク質に関して、高感度のLC-ESI-MS/MS systemを用いて解析を行った結果、Reference Sequence (NCBI)に登録されている52個のタンパク質(ORF：100アミノ酸残基以下)に加え、7個の新規タンパク質(同左)を同定することが出来た。この中で5つの新規タンパク質に関しては、相当する完全長cDNA配列中で既知ないし推定コード領域の上流に存在する短いORFがコードするタンパク質であった。この5つのタンパク質のATGコドンは全て各cDNA配列中で最も上流に存在しており、これらのタンパク質は典型的な翻訳開始のメカニズムに従って翻訳されたものと考えられる。多くの遺伝子の5'端の非翻訳領域に短いORFが存在することから、実際に細胞中で発現して機能を担っているタンパク質群の全体像を捕らえる上で、低分子タンパク質に関して更に解析を行うことが非常に重要であると考えられる。