テキストマイニングを用いた筋萎縮性側索硬化症の新規原因候補遺伝子の抽出

畠 野 雄 也
新潟大学大学院医歯学総合研究科
分子細胞医学専攻　脳神経内科学分野
（主任：小野寺　理 教授）

Identification of Novel Causative Genes for Amyotrophic Lateral Sclerosis Using Textmining

Yuya HATANO

Department of Neurology,
Niigata University Graduate School of Medical and Dental Science
(Director: Prof. Osamu ONODERA)

要　旨

【緒言】筋萎縮性側索硬化症（Amyotrophic lateral sclerosis: ALS）の原因遺伝子は30種類以上が知られている．しかし，大家系が少なくなる一方，エクソーム解析で遺伝子多型が多数同定され，これらの病的意義を検討するのが困難となっている．遺伝子多型を認める遺伝子の中から，ALSの原因としての候補遺伝子を効率よく抽出し，病的意義を解析することが求められている．テキストマイニングは，テキスト情報から，類似する性質の単語を抽出する方法である．近年，IBM社の人工知能Watsonを用い，テキストマイニングによりALSの原因候補遺伝子の抽出が行われた．しかし，得られた候補遺伝子に対して，当該遺伝子に変異を持つ例は見出されておらず，その意義は不明である．著者はオープンソースのテキストマイニングであるword2vec/fastTextを用いALSの原因候補遺伝子の抽出を試み，それらの遺伝子の変異の有無をALS剖検例で検討し，その意義を検証した．

【方法】word2vecおよびfastTextを用いて2000年から2019年にPubMed上に公開された英語論文163948報の抄録を解析した．ALSの原因遺伝子の約3割はRNA結合蛋白質（RNA-binding protein: RBP）である．よって，1164個のRBP遺伝子を，既知の11種類のALS原因RBP遺伝子との類似度（テキストマイニングにおけるcos類似度）が高い順に順位付けし，ALS原因遺伝子の候補順とした．上位10%をCST 10%遺伝子（cosine similarity top10% gene）と命名した．また，既報のメタアナライシスデータより，ALSと非神経疾患ないし一般人集団コントロール群の間で，Rare damaging variant（RDV）の出現頻度が，コントロールに比してALS群に有意に多い遺伝子をALS-RVgと定義した．本法の正当性を，CST 10%遺伝子の中で1）ALS-RVgの頻度，2）アノテーションに用いられたGene ontology（GO）termの解析，3）アミノ酸配列から予測した天然変性領域とcos類似度の順位との関連の検討を行った．

Reprint requests to: Yuya HATANO
Department of Neurology,
Brain Research Institute, Niigata University,
1-757 Asahimachi-dori, Chuo-ku,
Niigata 951-8585, Japan.

別刷請求先：〒951-8585　新潟市中央区旭町通1-757
新潟大学脳研究所脳神経内科
畠 野 雄 也

さらにCST 10%遺伝子のうちALS-RVgを新規ALS原因候補遺伝子とし，病理診断ALS108症例でRDVを検索した．
【結果】2012年度までのテキストデータを用いた解析で，word2vecではcos類似度上位10%内に，2013年以降に同定された原因遺伝子が含まれた．word2vecでは，CST 10%遺伝子群はそれ以外の遺伝子群と比して，ALS-RVgが有意に多かった（word2vec: p = 0.0044, fastText: p = 0.073）．しかし，多重検定補正後のq値 0.05未満の遺伝子はなかった．またGO解析では，CST 10%遺伝子は，スプライシングに関係する遺伝子が多かった．cos類似度上位上位の遺伝子群は下位の遺伝子群と比べて，天然変性領域の占める割合が大きい傾向にあった（word2vec: p < 0.0001, fastText: p < 0.0001）．新規ALS原因候補遺伝子として，9遺伝子が同定され，ALS108例中8例にRDVを認めた．下位10%のALS-RVgは2遺伝子のみで，RDVは認めず，RDVを持つ症例の比率に有意な差を認めた（p = 0.0039）．
【考察】CST 10%遺伝子ではそれ以下の群と比較して，ALS-RVgが有意に多かった．さらにCST 10%遺伝子群は，既存のALS原因遺伝子の特徴を有した．このことはテキストマイニングによる候補遺伝子抽出が機能している可能性を示した．新規ALS原因候補遺伝子にはスプライシング関連遺伝子4遺伝子にRDVを見いだした．これらRDVの病原性を今後検討したい．

キーワード：筋萎縮性側索硬化症，テキストマイニング，word2vec，fastText，エクソーム解析

緒　　言

筋萎縮性側索硬化症（Amyotrophic lateral sclerosis: ALS）は致死的な運動神経変性疾患である．ALSの約10%は，家族性ALS（Familial ALS: FALS）であり，30種類以上の原因遺伝子が知られている[1)2)]．また孤発性ALS（Sporadic ALS: SALS）の10%にもALSの原因遺伝子の変異が見いだされている[1)]．原因遺伝子の同定は病態機序の解明や，治療法開発に有用である[3)]．しかし，FALSにおける遺伝子変異の同定率は60%にとどまり，従来の遺伝子同定法の限界が示唆されている．

ALSの原因遺伝子の同定は，家系での連鎖解析法が用いられてきた[4)5)]．しかし，大家系や，血族婚家系が減少し，このような解析は困難となっている．また，病態機序から候補遺伝子を推定し変異の有無を検索する候補遺伝子アプローチも有用であった．SALSおよび大部分のFALSの神経細胞質内封入体の主要構成成分であるTDP-43（TAR DNA-binding protein of 43 kDa）を発現する*TARDBP*遺伝子の変異のALS症例での同定はその一例である[6)]．一方，SALSを含む多数例のエクソーム解析により，*TBK1*，*ANXA11*などが同定された[7)8)]．しかし，エクソーム解析により見出された多数の遺伝子多型について，各々の病原性の検証は困難である．

これらを打開するために，人工知能を用い，既知の原因遺伝子の特性から候補遺伝子を推定する手法が期待されている[2)]．ALSの原因遺伝子の中で*TARDBP*，*FUS*，*ATXN2*，*ANG*，*SETX*，*hnRNPA2B1*，*hnRNPA1*，*TAF15*，*GLE1*，*MATR3*，*ARHGEF28*の11遺伝子はRNA結合蛋白質（RNA-binding protein: RBP）である[2)]．Liらは，この性質と，病的な遺伝子がプリオンドメインを持つことから，プリオンドメインへの類似性を計算し，複数のRBPを候補遺伝子として推定した[9)]．しかし，RBPは1500種以上存在し，そのいずれがより可能性が高いか効率よく推定するのは困難である[10)]．

近年，人工知能を用いて大量のテキストデータから有用情報を抽出する方法として，テキストマイニングが開発された．本法は，機能的ゲノミクス，バイオロジカルパスウェイ，タンパク質-タンパク質相互作用，薬物-遺伝子関連，比較トキシコゲノミクス，神経精神疾患領域の病態解明な

ど，様々な生物学的問題の研究に応用されている[11]．ALS 関連遺伝子についても，IBM 社の人工知能「Watson」を用い，学術論文を解析し，既知の ALS 原因遺伝子と共通した特性をもつ遺伝子群がテキストマイニングにより推定された[2]．しかし，それら遺伝子に多型をもつ患者が同定されておらず，病原性は不明である．

テキストマイニングを行えるツールとして，高価な Watson 以外にも，オープンソースの解析ツールとして，word2vec（https://github.com/svn2github/word2vec）[12] と fastText（https://github.com/facebookresearch/fastText）[13] が知られている．word2vec は Google 社にて開発された自然言語処理ツールである．自然言語処理とは人間の言語（自然言語）を統計的に解析できる形に変換し機械処理を行うことであり，単語を数値ベクトル化して互いの意味関係を定量化する．fastText は Word2vec を発展させたツールであり，より解析速度に優れ出現頻度の少ない単語でも対応できるという特徴がある[13]．

著者はこれらのオープンソースを用い ALS 関連論文をテキストマイニングすることにより，ALS の新たな原因遺伝子の候補を抽出し，それら遺伝子変異が ALS 病理例で濃縮されているか検証した．

方　　法

テキストマイニング概要

解析対象とした英語論文抄録から，自然言語処理において性能に悪影響を与える，ストップワード（「the」「is」など）[14] 及び '[', ']', '#', '&', ',', '.', ';', ':', ' (', ') ', '%', '<', '>', '!', '?' の記号を削除し，解析テキストとした（**図 1 A**）．ストップワードには ntlk パッケージの stopwords を用いた[14]．オープンソース解析ツールとして，word2vec（https://github.com/svn2github/word2vec）[12] と fastText（https://github.com/facebookresearch/fastText）[13] を使用した．数値ベクトル化の手法としては，周辺語から中心の単語を予測する CBOW と，中心の単語から周辺語を予測する skipgram とがある（**図 1 B**）．word2vec はデフォルト設定である CBOW，fastText は精度が高い skipgram で解析した[15]．単語間の類似性は解析によって得られた数値ベクトル同士の角度の近さ（cos 類似度）で評価した（**図 1 C, D**）[16]．ウィンドウサイズ（周辺語として中心の単語の前後何単語まで考慮するか）はデフォルト設定の 5 とした．ベクトルの次元数は word2vec では 既報[17] を参考に 200 とし，fastText ではデフォルト設定の 100 とした．fastText では稀な単語に対応するために，単語を n 文字毎に区切りベクトル付けを行い，そのパラメーターを n-grams と表現する．本解析では n=1 と設定した[18]．

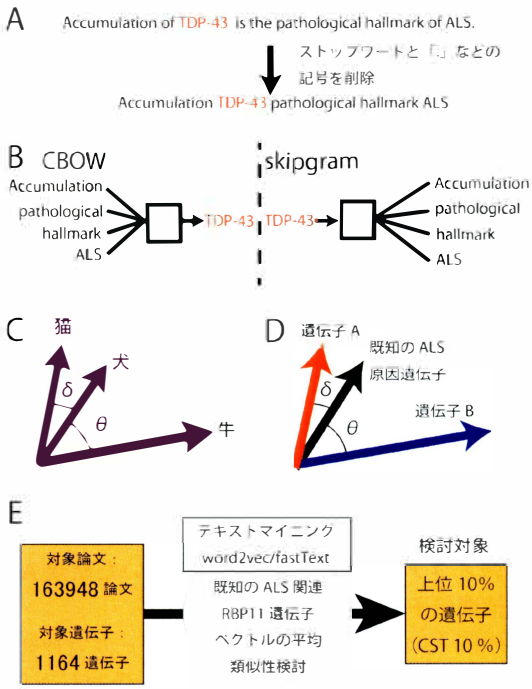

図 1　（A）ストップワードと記号を削除して解析テキストとした．（B）CBOW は周辺語から中心語を予測する．Skipgram は中心語から周辺語を予測する．（C）各単語とのベクトルを互いになす角の cos で比較する．図中では犬は牛より猫に類似している．（D）既知の ALS 原因遺伝子と候補遺伝子の類似性を cos で比較する．（E）word2vec/fastText を用いた解析の概略図

テキストマイニングによる解析方法

「Watson」を用いた既報を参考におこなった[2]. 具体的には，PubMed（https://pubmed.ncbi.nlm.nih.gov/）の英語論文にて，抄録に”RNA-binding protein”を含む論文を抽出した．解析対象とする遺伝子は，既報[2]でRBP遺伝子として定義され，かつ抽出した抄録に記載のある遺伝子とした．word2vecまたはfastTextにて既知ALS関連RBP遺伝子のベクトルの平均と，各々の遺伝子のcos類似度を比較し，その順位を評価した（**図1E**）．また抄録中に一度でも使用された単語はすべて解析対象とした．本稿では，cos類似度が解析遺伝子中，上位10%以内の遺伝子をCST 10%遺伝子（cosine similarity top10% gene）とした．

既報メタアナライシスデータを用いた新規ALS候補遺伝子の検証

中国人のALS 610例とコントロール（神経疾患の家族歴や病歴がなく，病院を受診した例）460例およびヨーロッパ人のALS 2876例とコントロール（HudsonAlpha社，McGill大学の協力で得られた一般人集団群．神経疾患のスクリーニングは行われていない．）6405例で遺伝子のRare damaging variant（RDV）をCochran-Mantel-Haenszel test（CMH）を用いて比較した既報のメタアナライシスデータを利用した[7,19]．この既報におけるRDVの定義は，内部データでのアレル頻度が5×10^{-4}未満かつ遺伝子変異データベースExac（https://gnomad.broadinstitute.org/）でのアレル頻度が5×10^{-5}未満の非同義のvariantであり，かつ，変異による蛋白毒性度をPolyphen2（http://genetics.bwh.harvard.edu/pph2/）にて予測し良性と判断されたものを除いたvariantである．このvariantデータを基にして，既に既報で解析されている，各々の遺伝子のCMH p valueを利用した[19]．

本稿では，CMH p value < 0.05未満であり，多重検定補正前の検討で，ALSとコントロール群とで有意にRDVの出現頻度がALS群に多かった遺伝子をALS-RVgと定義し，CST 10 %遺伝子のうちALS-RVgのものを新規ALS原因候補遺伝子とした．

天然変性領域の予測

各遺伝子由来蛋白質について，アミノ酸配列から蛋白質の天然変性領域をIUPred2A（https://iupred2a.elte.hu/）[20]にて予測した．Score 0.5をカットオフ値として，それを超える値を天然変性領域とし，アミノ酸配列全体に占める推定天然変性領域の割合を求めた．

病理診断ALS例での遺伝情報解析

1978年～2019年に新潟大学脳研究所病理学教室で剖検され，TDP-43陽性の封入体を認め，病理学的にALSと診断したSALS 139例を対象とした．SALSは，二親等以内に病歴からALS発症者を認めない症例とした．

ゲノムDNAは，ALS患者では剖検組織の中枢神経系（後頭葉，運動皮質，小脳），コントロールでは血液検体から，DNA抽出キット（QIAamp DNA Mini Kit; Bio-Rad Laboratories, Hercules, CA, USA）を使用して抽出した．DNA品質についてはAgilent 2200 tape station（Agilent Technologies, Santa Clara, USA）で確認した．エクソーム解析はライブラリ作成，シークエンス解析，アノテーション解析の工程はタカラバイオ社にて，Illumina NovaSeq 6000にて施行された．

アノテーション解析データを用いて，新規ALS原因候補遺伝子の変異を検出した．よく知られた病的意義を持つ一塩基変異である*SOD1* p.D91AのアレルgnomADの全人種でのデータベースで0.001であることから[21]，アレル頻度が0.001未満の変異を解析対象とした．さらに非同義の一塩基変異の中で，Polyphen2（http://genetics.bwh.harvard.edu/pph2/）で良性と判断されたものを除いたものをRare damaging variant（RDV）とした．アレル頻度のデータベースとしては，Human Genetic Variation Database（HGVD）（http://www.hgvd.genome.med.kyoto-u.ac.jp/）とExome Aggregation Consortium All / East Asia（ExAC All/EAS）（https://gnomad.broadinstitute.org/）を用いた．

統計解析

CST 10% 遺伝子は非 CST10％遺伝子と比べて ALS-RVg が多いかを Fisher の正確確率検定を用いて両側検定を行った．ALS と一般人集団コントロールでの遺伝子で RDV を比較した既報の Cochran-Mantel-Haenszel test について，CST 10% 遺伝子で，多重検定補正を Benjamini-Hochberg 法を用いて行い，q 値を算出した．また http://geneontology.org/（release 2021-05-01）に基づいて，CST 10 ％遺伝子および既報[2]で RBP と定義された 1164 遺伝子のうち特定の Gene ontology（GO）でアノテーションされた遺伝子の数を求めた．さらに CST 10 ％遺伝子と RBP 遺伝子全体の比較でその割合について，Fisher の正確確率検定を用いて，両側検定を行った．多重検定補正は Benjamini-Hochberg 法を用いた．Word2vec/fastText での cos 類似度の順位を 10％ごとに階層化し，上位の群ほど天然変性領域の予測割合が大きいか，Jonckheere-Terpstra trend test で検定した．CST 10 ％遺伝子と下位 10％の遺伝子の ALS-RVg の RDV の頻度に有意差があるか，母比率の差の検定を用いて検定した．いずれの検定も有意水準は 5％とした．統計ソフトとして，R version 4.1.0 を使用した．

結　　果

既知の ALS 原因遺伝子の同定によるテキストマイニングの妥当性の検証

まず，Word2vec および fastText を用いたテキストマイニングにて ALS 原因遺伝子を抽出できるか検証した[2]．具体的には，2000 年 1 月 1 日から 2012 年 12 月 31 日に PubMed に公開された英語論文から，2012 年までに同定された，ALS 関連 RBP 遺伝子 *TARDBP, FUS, ATXN2, ANG, SETX, hnRNPA2B1, hnRNPA1, TAF15* のベクトル平均を word2vec または fastText で算出した．この既知遺伝子ベクトル平均との cos 類似度を，「Watson」を用いた既報で解析されていた 878 個の RBP 遺伝子それぞれについて算出し[2]，順位付けを行った．次に，2013 年以降に同定された ALS の原因遺伝子で RBP である *ARHGEF28, MATR3, GLE1* 遺伝子について，878 個の RBP 遺伝子内での順位を算出した．Word2vec では *MATR3, GLE1* は，各々 34 位と 56 位と上位 10％に抽出された．fastText では各々 241 位と 671 位と 27.4% と 76.4% であった．*ARHGEF28* は抄録中に一度も使用されておらず，評価できなかった．

テキストマイニングによる新規 ALS 原因候補遺伝子の抽出

2000 年 1 月 1 日から 2019 年 8 月 27 日に PubMed に公開された英語論文を用い，既知 ALS 関連 RBP 遺伝子 *TARDBP, FUS, ATXN2, ANG, SETX, hnRNPA2B1, hnRNPA1, TAF15, GLE1, MATR3, ARHGEF28* のベクトル平均を算出した．次に，「Watson」を用いた既報で解析された 1164 の遺伝子と[2]，先の ALS 関連 RBP 遺伝子のベクトル平均との cos 類似度を算出し，これが高い順に順位付けた（**図 1 E**）．

この方法の正当性について，上位にランクされた遺伝子の RDV が，ALS 群に多いか否かを，既報のメタアナライシスデータを用い検証した[7)19)]．2012 年までの word2vec による解析で ALS 原因 2 遺伝子が，cos 類似度上位 10％以内の遺伝子（CST 10％遺伝子）に含まれたことから，本解析もこの範囲に着目した．

まず CST 10 ％遺伝子の遺伝子について GO term 解析をおこない，その機能を検討した．その結果 word2vec 抽出遺伝子群ではスプライシングに関わる遺伝子を多く認め（**図 2 A**），fastText 抽出遺伝子群ではスプライシングと核内小体に関わる遺伝子を多く認めた．（**図 2 B**）．さらにこれら抽出遺伝子由来の蛋白質において，天然変性領域の占める割合を検討した．既知の ALS 関連 RBP 遺伝子は，その構造中に天然変性領域の占める割合が高い特徴があり平均で 51.3% であった．CST10% の遺伝子群において，天然変性領域の占める割合は 47.0%（word2vec）と 43.8 %（fastText）と高値であった．一方，下位 10％遺

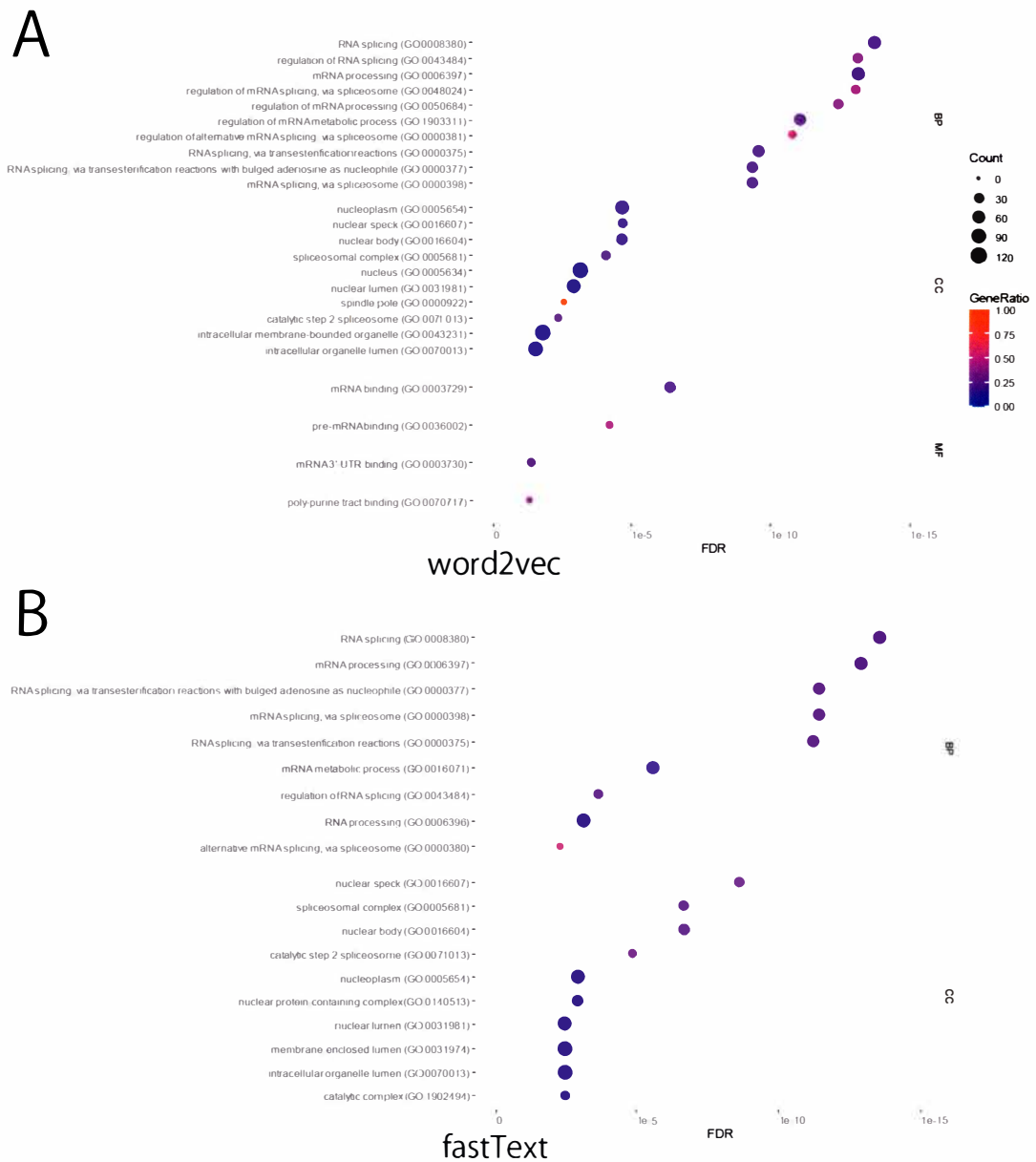

図 2　(A) word2vec と (B) fastText で解析を行い，RBP 遺伝子 1164 遺伝子との比較で CST 10% 遺伝子に多い GO term を dot plot で表した．Biological process (BP)，Cellular compornent (CC)，Molecular function (MF) の各項目で FDR (false discovery rate) が上位 10 位以内で 0.05 未満の GO term をプロットした．点の大きさは特定の GO term が上位 10% の遺伝子に認められた数を表す (Count)．GeneRatio は特定の GO term が上位 10% の遺伝子に認められた数 / RBP 遺伝子 1164 遺伝子に認められた数を表す．

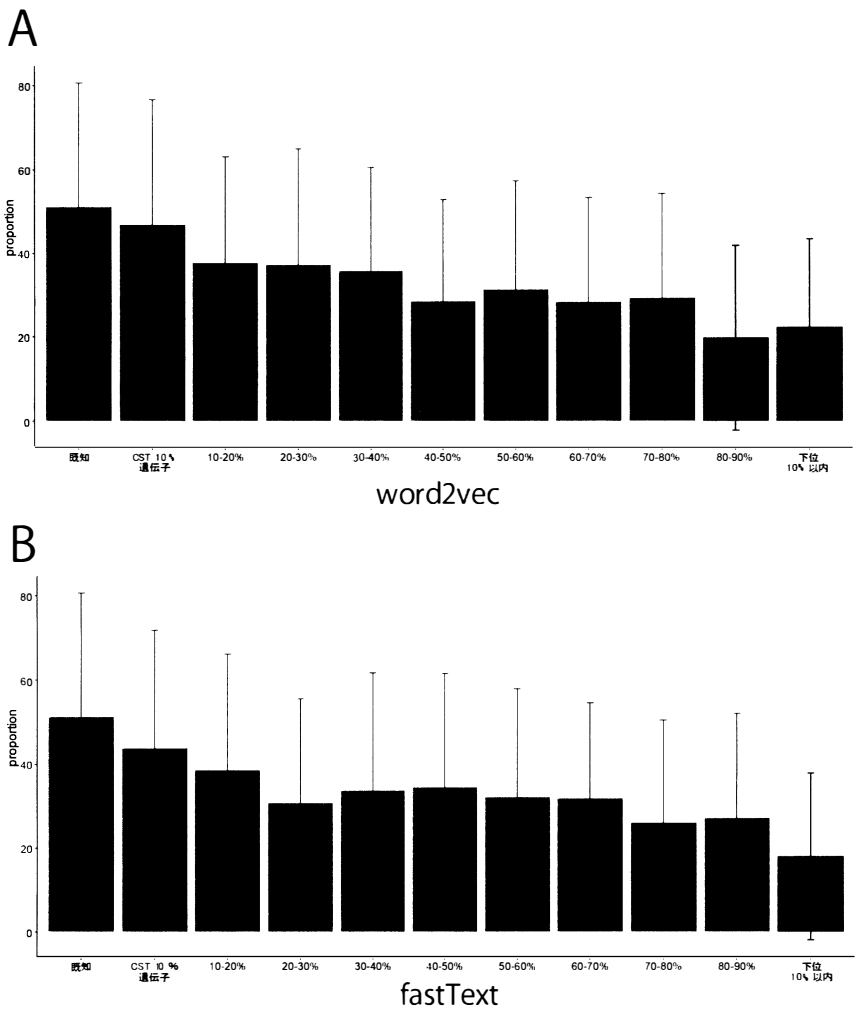

図3　cos 類似度の順位を 10% ごとに層別化して RBP 遺伝子を割り当てて，各層ごとの遺伝子の IUPred2A で予測した天然変性領域の割合の平均を棒グラフに示した．エラーバーは標準偏差である．(A) が word2vec，(B) が fastText での解析である．「既知」は既知の ALS 原因遺伝子である．

伝子では 22.4% (word2vec) と 18.1% (fastText) と低く，順位が高い層ほど予測される天然変性領域の割合が高くなる傾向を認めた (word2vec: $p < 0.0001$, fastText: $p < 0.0001$, **図3 A, B**)．

次に RDV がコントロールより ALS 群にて高頻度に認められる遺伝子を ALS-RVg とし，その出現頻度を cos 類似度の順位を 10% 毎に階層化し比較した (**図4 A, B**)．ALS-RVg の割合は，CST 10% 遺伝子において word2vec と fastText でそれぞれ 10.3% (7/68) と 7.1% (5/70)，同 10% 未満の遺伝子では 2.6% (16/624) と 2.9% (18/622) であり，word2vec では有意に CST 10% 遺伝子群で多かった (word2vec: $p = 0.0044$, fastText: $p = 0.073$) (**図4 A, B**)．quantile-quantile plot (Q-Q plot) では，word2vec と fastText のいずれも，CST 10 %遺伝子では，ALS が関連している遺伝子がないと仮定した場合の期待値の p 値より観測値の p 値が小さい傾向にあった (**図4 C, D**)．

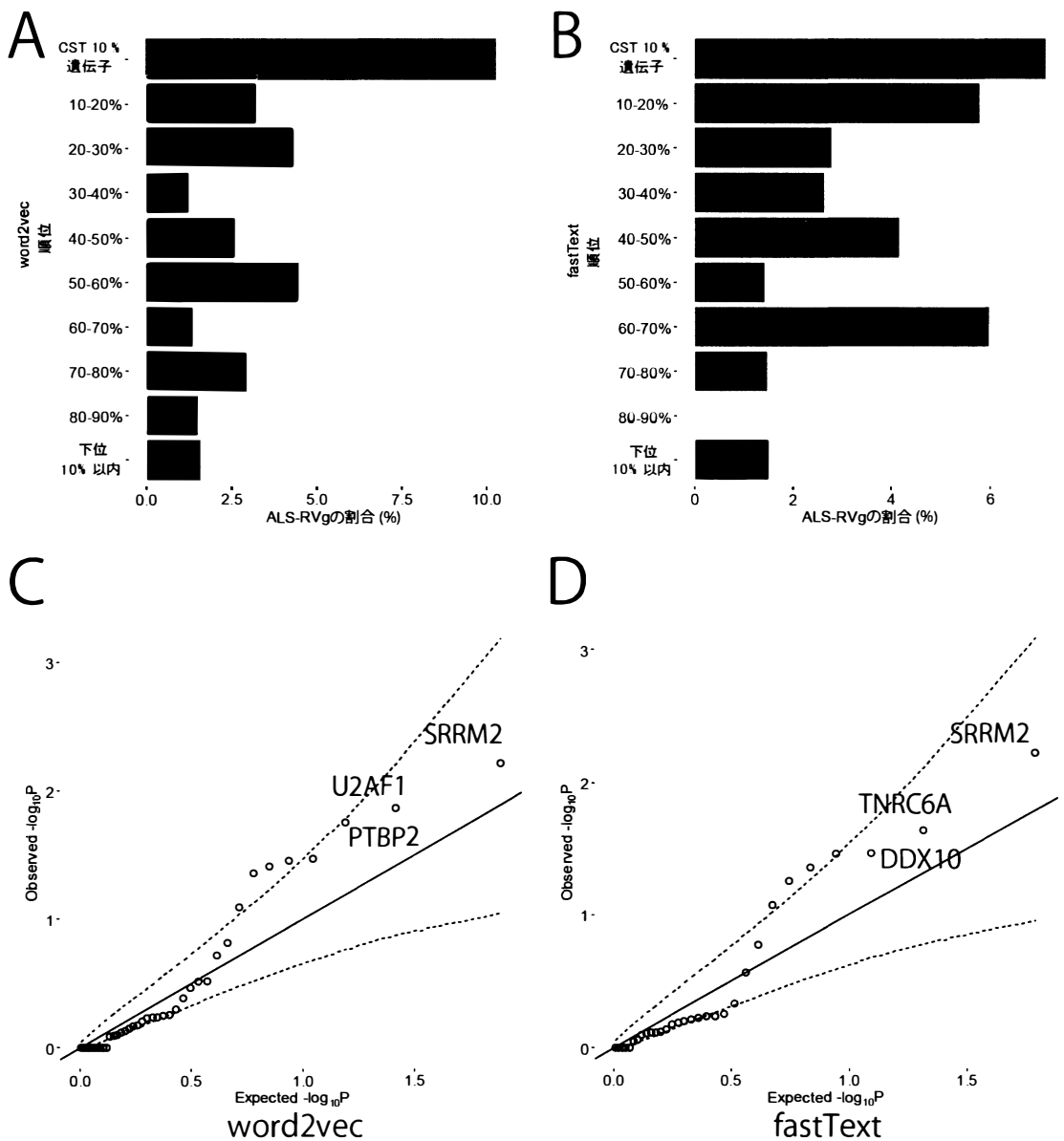

図4　(A, B) cos類似度の順位を10%ごとに層別化して，RBP遺伝子を割り当てて，各層ごとのメタアナライシスデータで解析されている遺伝子のうちALSとcontrol群でRVの出現頻度が有意にALS群に多かった遺伝子(ALS-RVg)の割合．(A)がword2vec，(B)がfastTextでの解析．(C, D) quantile-quantile plot (Q-Q plot)．CST 10%遺伝子におけるメタアナライシスデータのRV頻度がALS/Controlでで ALS群の方が多い遺伝子のうちCMHで観測されたp値とp値の期待値の関係を表す．点線の範囲内が95%信頼区間を表す．(C)がword2vec，(D)がfastTextでの解析である．

表1　ALS関連候補遺伝子

遺伝子	p値	q値	OR	天然変性領域割合(%)
word2vec				
SRRM2	0.006033	0.20211	1.37[1.09-1.72]	97.5
U2AF1	0.013735	0.295171	-[1.25-]	23.3
PTBP2	0.017622	0.295171	-[1.14-]	12.1
RBM20	0.033782	0.366336	1.89[1.03-3.47]	69.6
DHX9	0.034723	0.366336	2.76[1.01-8.01]	13.9
SON	0.038274	0.366336	1.46[1.00-2.10]	68.2
EXOSC3	0.044051	0.36893	5.47[0.82-60.95]	7.64
fastText				
SRRM2	0.006033	0.21116	1.37[1.09-1.72]	97.5
TNRC6A	0.022874	0.270068	1.66[1.06-2.60]	86.9
DDX10	0.034289	0.270068	1.85[1.01-3.34]	39.8
DHX9	0.034723	0.270068	2.76[1.01-8.01]	13.9
EXOSC3	0.044051	0.308359	5.47[0.82-60.95]	7.64

表2　RDVおよびRDVを認めた症例の情報

Case	gene	Amino acid change	genetic database			sex	Age of onset(y)	survival time(m)	dementia	Symptoms at onset	TDP-43/FUS pathology
			HGVD	ExAC All	ExAC EAS						
Case 1	SRRM2	p.Arg33Trp	-	8.80E-06	0	F	45	72	unknown	lower limbs	TDP-43
Case 2	SRRM2	p.Pro2073Ser	-	-	-	F	53.6	11	-	upper limbs	TDP-43
Case 3	RBM20	p.Ile536Thr	0.0004	-	-	M	63	26	-	upper limbs	TDP-43
Case 4	RBM20	p.Gly697Arg	0.0008	-	-	M	63.4	40	-	upper limbs	TDP-43
Case 5	RBM20	p.Gly1217Arg	0.0008	-	-	F	73.25	90	+	dementia	TDP-43
Case 6	TNRC6A	p.Trp1064Ser	-	-	-	F	72	20	+	dementia	TDP-43
Case 7	TNRC6A	p.Arg1363His	-	1.65E-05	0	M	59	24	unknown	upper limbs	TDP-43
Case 8	SON	p.Pro397Arg	-	-	-	F	64.3	27	unknown	bulbar + upper limbs	TDP-43

Abbreviations: HGVD, Human Genetic Variation Database; ExAC All, Exome Aggregation Consortium (All population); ExAC EAS, Exome Aggregation Consortium (East Asia population); y, year; m, month; F, female; M, male.

これらCST 10 %遺伝子かつALS-RVgの9遺伝子を新規ALS原因候補遺伝子とした（**表1**）．9遺伝子のうち，多重検定補正後のq値で有意差を認めた遺伝子は存在しなかった（**表1**）．

ALS原因候補遺伝子の変異の検討

word2vecおよびfastTextで見いだされた新規ALS原因候補遺伝子9遺伝子のRDVを，病理診断されたSALS 139例のうち，既知の遺伝子に変異が見いだされた例を除外した108例にて検討した．108例中8例（7.4%）に新規ALS原因候補遺伝子のRDVを見いだした．内訳は，両解析で共通して抽出された*SRRM2*（2例），word2vecのみで抽出された*RBM20*（3例），*SON*（1例），fastTextのみで抽出された*TNRC6A*（2例）であった（**表2**）．一方，下位10%以内の遺伝子では，ALS-RVgは2遺伝子のみで，これらについては108例中にRDVを認めなかった．新規ALS原因候補遺伝子と下位10%遺伝子由来のALS-RVgでは，RDVを持つ症例の比率に有意な差を認めた（$p = 0.0039$）．

考　察

著者はオープンソースのテキストマイニングを用い，ALSの新規原因候補遺伝子の抽出を行った．テキストマイニングはパラメーター設定により得られる結果が変動しうる．よって2種類のword2vecおよびfastTextを用いた．2012年度までのテキストデータを用い，2013年以降に同定された遺伝子が抽出できるか解析した．その結果，word2vecではcos類似度上位10%内に，原因遺伝子が含まれた．一方fastTextでは上位に抽出し得なかった．今回のALSのRBPを対象とした解析では，word2vecの方が，目的とする遺伝子を抽出しやすい可能性が考えられた．

その結果より，2019年までのテキストデータ解析にてcos類似度上位（CST）10%に抽出された遺伝子を解析対象とした．さらに，そのうち既報のメタアナライシスデータ[7)19)]でALS群にRDVが多く認められた9遺伝子をALS原因候補遺伝子とした．病理診断したSALS患者108例でのエクソーム解析では，同候補遺伝子のRDVを8例に見いだした．一方下位10%遺伝子では，108例中にRDVは検出し得なかった．この結果は，先の8個のRDVの病原性を示唆すると考えた．今後はこれらの症例の病理解析，変異の分子生物学的解析を進め，これらの変異の病的意義を検討し，ALSとの関連を立証する必要がある．

今回のテキストマイニングにてALS関連遺伝子が上位遺伝子群に濃縮されている可能性は，次の結果にて示唆される．まず，word2vecでは，CST 10%遺伝子群で，ALS-RVgの頻度が，それ以下の順位の遺伝子群よりも有意に多かった．実際，Q-Q plotでも，CST 10%遺伝子では，ALS関連遺伝子がないと仮定した場合に期待されるp値の分布より観測値のp値が小さい傾向にあった．さらに，上位にリストされた遺伝子は，ALS原因遺伝子の特性を有していた．上位の遺伝子由来蛋白質は天然変性領域が占める割合が大きい傾向にあった（**図3**）．天然変性領域は，液相分離現象（Liquid liquid phase separation: LLPS）を介して，凝集体形成をきたし，ALS発症に関与する[22)23)]．実際，既知のALS原因RBP遺伝子の多くは天然変性領域/プリオン様ドメインを有している[9)]．この結果も，本法がALS関連遺伝子を濃縮していることを示唆する．最後に，ALS病態生理の一つにスプライシング異常が推定されているが，GO解析の結果，CST 10%遺伝子にはスプライシング関連遺伝子が濃縮されていた．

今回見いだした*SRRM2, RBM20, TNRC6A, SON*は，いずれもスプライシング関連遺伝子である[24)-27)]．ALS患者組織では，スプライシング異常や[28)]，mRNAのスプライシングに関与する機能性RNA snRNAの低下が報告されている[29)]．また，ALSの神経細胞内封入体の主要構成成分であるTDP-43はスプライシングによって量が自己調節されており，ALSではその自己調節機構が破綻している[30)]．今後は今回見いだされた変異がこれらの遺伝子の機能異常を介して，スプライシングに影響するか否かを，培養細胞系や罹患組織で検討したい．

本研究の限界点について述べる．まず，本解析で用いた病理診断例エクソーム解析データの症例数の問題がある．病理診断例であるため，確実な診断例であり正確な予後評価も実施できる利点は大きいが，一方で原因遺伝子同定の検証のためには症例数が十分でない．これについては，今後，他施設とも協力してより大規模な検討を行いたい．また今回見いだされた原因候補遺伝子やその変異の病的意義を検討するためには，機能解析が必須であるが，その機能解析のスタンダードがなく病原性の立証が困難である．また，今回の検討ではfastTextはword2vecほどALS関連遺伝子抽出に有効でなかった可能性がある．fastTextでは，2012年までの学習で既知の遺伝子を上位に抽出できず，2019年までの学習ではCST 10%遺伝子群でALS-RVgが有意な濃縮に至らなかった．この結果については，単語を区切ってベクトル付けを行うというfastTextの特性が災いしている可能性がある．この特性は，語形変化が多い言語には有効と考えられるが，遺伝子名のよう

に表記が似ていても異なるものを表す場合は，誤った関連付けを行う可能性がある[31]．fastText解析でのより適切なパラメーター設定についても検討する必要がある．

致死的疾患であるALSの治療法開発には，病態解明に加えて質の高い臨床治験が必要となる．ALSの遺伝的背景は多彩であり，その正確な診断は臨床治験におけるグループ分け，結果解析に重要となる可能性がある．テキストマイニングを用いることで，候補遺伝子の効率的な抽出が可能となりうる．その病原性の有無を検討し，見いだされた変異の病的意義を検証とすることにより，本症の正確な診断と，治験における層別化を可能とし，ALS病態機序の解明と治療薬開発に貢献することが期待される．

謝　辞

本研究にあたり貴重なテーマをご指導いただきました新潟大学脳研究所脳神経内科　教授　小野寺理先生，研究全般にわたり直接的な指導をいただきました新潟大学脳研究所脳神経内科　石原智彦先生，ALS剖検検体を使用させていただきました新潟大学脳研究所病態神経科学部門病理学教室　教授　柿田明美先生，他田真理先生，貴重なご意見をいただきました新潟大学大学院医歯学総合研究科バイオインフォマーティクス分野　奥田修二郎先生，その他ご指導，ご支援を頂いた諸先生方に深謝いたします．

参考文献

1）Chia R, Chiò A, and Traynor BJ: Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. 17: 94-102, 2018.

2）Bakkar N, Kovalik T, Lorenzini I, Spangler S, Lacoste A, Sponaugle K, Ferrante P, Argentinis E, Sattler R, and Bowser R: Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. Acta Neuropathol. 135: 227-247, 2018.

3）Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC and Sanseau P: The support of human genetic evidence for approved drug indications. Nat Genet. 47: 856 60, 2015.

4）Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng HX, Rahmani Z, Krizus A, McKenna Yasek D, Cayabyab A, Gaston SM, Berger R, Tanzi RE, Halperin JJ, Herzfeldt B, den Bergh RV, Hung WY, Bird T, Deng G, Mulder DW, Smyth C, Laing NG, Soriano E, Pericak Vance MA, Haines J, Rouleau GA, Gusella JS, Horvitz HR, and Brown RH Jr: Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature. 362: 59-62, 1993.

5）Maruyama H, Morino H, Ito H, Izumi Y, Kato H, Watanabe Y, Kinoshita Y, Kamada M, Nodera H, Suzuki H, Komure O, Matsuura S, Kobatake K, Morimoto N, Abe K, Suzuki N, Aoki M, Kawata A, Hirai T, Kato T, Ogasawara K, Hirano A, Takumi T, Kusaka H, Hagiwara K, Kaji R, and Kawakami H: Mutations of optineurin in amyotrophic lateral sclerosis. Nature. 465: 223-226, 2010.

6）Yokoseki A, Shiga A, Tan CF, Tagawa A, Kaneko H, Koyama A, Eguchi H Tsujino A, Ikeuchi T, Kakita A, Okamoto K, Nishizawa M, Takahashi H, and Onodera O: TDP-43 mutation in familial amyotrophic lateral sclerosis. Ann Neurol. 63: 538-542, 2008.

7）Cirulli ET, Lasseigne BN, Petrovski S, Sapp PC, Dion PA, Leblond CS, Couthouis J, Lu YF, Wang Q, Krueger BJ, Ren Z, Keebler J, Han Y, Levy SE, Boone BE, Wimbish JR, Waite LL, Jones AL, Carulli JP, Day-Williams AG, Staropoli JF, Xin WW, Chesi A, Raphael AR, McKenna-Yasek D, Cady J, Vianney de Jong JM, Kenna KP, Smith BN, Topp S, Miller J, Gkazi A; FALS Sequencing Consortium, Al-Chalabi A, van den Berg LH, Veldink J, Silani V, Ticozzi N, Shaw CE, Baloh RH, Appel S, Simpson E, Lagier-Tourenne C, Pulst SM,

Gibson S, Trojanowski JQ, Elman L, McCluskey L, Grossman M, Shneider NA, Chung WK, Ravits JM, Glass JD, Sims KB, Van Deerlin VM, Maniatis T, Hayes SD, Ordureau A, Swarup S, Landers J, Baas F, Allen AS, Bedlack RS, Harper JW, Gitler AD, Rouleau GA, Brown R, Harms MB, Cooper GM, Harris T, Myers RM and Goldstein DB: Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. 347: 1436-1441, 2015.

8) Smith BN, Topp SD, Fallini C, Shibata H, Chen HJ, Troakes C, King A, Ticozzi N, Kenna KP, Soragia-Gkazi A, Miller JW, Sato A, Dias DM, Jeon M, Vance C, Wong CH, de Majo M, Kattuah W, Mitchell JC, Scotter EL, Parkin NW, Sapp PC, Nolan M, Nestor PJ, Simpson M, Weale M, Lek M, Baas F, Vianney de Jong JM, Ten Asbroek ALMA, Redondo AG, Esteban-Pérez J, Tiloca C, Verde F, Duga S, Leigh N, Pall H, Morrison KE, Al-Chalabi A, Shaw PJ, Kirby J, Turner MR, Talbot K, Hardiman O, Glass JD, De Belleroche J, Maki M, Moss SE, Miller C, Gellera C, Ratti A, Al-Sarraj S, Brown RH Jr, Silani V, Landers JE and Shaw CE: Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. Sci Transl Med. 9: eaad9157, 2017.

9) Li YR, King OD, Shorter J and Gitler AD: Stress granules as crucibles of ALS pathogenesis. J Cell Biol. 201: 361-372, 2013.

10) Gerstberger S, Hafner M and Tuschl T: A census of human RNA-binding proteins. Nat Rev Genet. 15: 829-845, 2014.

11) Abbe A, Grouin C, Zweigenbaum P and Falissard B: Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. 25: 86-100, 2016.

12) Mikolov T, Corrado G, Chen K and Dean J: Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/pdf/1301.3781v3.pdf, 2013.

13) Bojanowski P, Grave E, Joulin A and Mikolov T: Enriching Word Vectors with Subword Information. Preprint at https://arxiv.org/pdf/1607.04606.pdf, 2017

14) Cheng MY, Kusoemoro D and Gosno RA: Text mining-based construction site accident classification using hybrid supervised machine learning. Automation in Construction. 118: 103265, 2020.

15) Gokul S Krishnan and Sowmya Kamath S: Evaluating the quality of word representation models for unstructured clinical Text based ICU mortality prediction. ICDCN '19: Proceedings of the 20th International Conference on Distributed Computing and Networking. 480-485, 2019.

16) Orkphol K and Yang W: Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. 11: 114, 2019.

17) Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson KA, Ceder G and Jain A: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 571: 95-98, 2019.

18) https://github.com/facebookresearch/fastText/issues/585

19) Gratten J, Zhao Q, Benyamin B, Garton F, He J, Leo PJ, Mangelsdorf M, Anderson L, Zhang ZH, Chen L, Chen XD, Cremin K, Deng HW, Edson J, Han YY, Harris J, Henders AK, Jin ZB, Li Z, Lin Y, Liu X, Marshall M, Mowry BJ, Ran S, Reutens DC, Song S, Tan LJ, Tang L, Wallace RH, Wheeler L, Wu J, Yang J, Xu H, Visscher PM, Bartlett PF, Brown MA, Wray NR and Fan D: Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. Genome Med. 9: 97, 2017.

20) Mészáros B, Erdos G and Dosztányi Z: IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 46: W329-W337, 2018.

21) Dols-Icardo O, García-Redondo A, Rojas-García

R, Borrego-Hernández D, Illán-Gala I, Muñoz-Blanco JL, Rábano A, Cervera-Carles L, Juárez-Rufián A, Spataro N, De Luna N, Galán L, Cortes-Vicente E, Fortea J, Blesa R, Grau-Rivera O, Lleó A, Esteban-Pérez J, Gelpi E and Clarimón J: Analysis of known amyotrophic lateral sclerosis and frontotemporal dementia genes reveals a substantial genetic burden in patients manifesting both diseases not carrying the C9orf72 expansion mutation. J Neurol Neurosurg Psychiatry. 89: 162-168, 2018.

22) Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J, Grishin NV, Frantz DE, Schneider JW, Chen S, Li L, Sawaya MR, Eisenberg D, Tycko R and McKnight SL: Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell. 149: 753-767, 2012.

23) Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynov S, Mahamid J, Saha S, Franzmann TM, Pozniakovski A, Poser I, Maghelli N, Royer LA, Weigert M, Myers EW, Grill S, Drechsel D, Hyman AA and Alberti S: A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell. 162: 1066-1077, 2015.

24) Tanaka H, Kondo K, Chen X, Homma H, Tagawa K, Kerever A, Aoki S, Saito T, Saido T, Muramatsu SI, Fujita K and Okazawa H: The intellectual disability gene PQBP1 rescues Alzheimer's disease pathology. Mol Psychiatry. 23: 2090-2110, 2018.

25) Maatz H, Jens M, Liss M, Schafer S, Heinig M, Kirchner M, Adami E, Rintisch C, Dauksaite V, Radke MH, Selbach M, Barton PJ, Cook SA, Rajewsky N, Gotthardt M, Landthaler M and Hubner N: RNA-binding protein RBM20 represses splicing to orchestrate cardiac pre-mRNA processing. J Clin Invest. 124: 3419-3430, 2014.

26) Suzawa M, Noguchi K, Nishi K, Kozuka-Hata H, Oyama M and Ui-Tei K: Comprehensive Identification of Nuclear and Cytoplasmic TNRC6A-Associating Proteins. J Mol Biol. 429: 3319-3333, 2017.

27) Sharma A, Markey M, Torres-Muñoz K, Varia S, Kadakia M, Bubulya A and Bubulya PA: Son maintains accurate splicing for a subset of human pre-mRNAs. J Cell Sci. 124: 4286-4298, 2011.

28) Shiga A, Ishihara T, Miyashita A, Kuwabara M, Kato T, Watanabe N, Yamahira A, Kondo C, Yokoseki A, Takahashi M, Kuwano R, Kakita A, Nishizawa M, Takahashi H and Onodera O: Alteration of POLDIP3 splicing associated with loss of function of TDP-43 in tissues affected with ALS. PLoS One. 7: e43120, 2012.

29) Ishihara T, Ariizumi Y, Shiga A, Kato T, Tan CF, Sato T, Miki Y, Yokoo M, Fujino T, Koyama A, Yokoseki A, Nishizawa M, Kakita A, Takahashi H and Onodera O: Decreased number of Gemini of coiled bodies and U12 snRNA level in amyotrophic lateral sclerosis. Hum Mol Genet. 22: 4136-4147, 2013.

30) Koyama A, Sugai A, Kato T, Ishihara T, Shiga A, Toyoshima Y, Koyama M, Konno T, Hirokawa S, Yokoseki A, Nishizawa M, Kakita A, Takahashi H and Onodera O: Increased cytoplasmic TARDBP mRNA in affected spinal motor neurons in ALS caused by abnormal autoregulation of TDP-43. Nucleic Acids Res. 44: 5820-5836, 2016.

31) 平出 聡, 田中瑛一, 大西健司：日本語の文字種を考慮した単語の分散表現の学習手法. 人工知能学会全国大会論文集 34: 1D5GS903, 2020.

（令和3年6月18日受付）