様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　１　日現在

機関番号：　３２６６０
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０３５０

研究課題名（和文）ノンコーディング遺伝子の発現制御領域の数理的解析

研究課題名（英文）Computational analysis for the control regions of non-ccding genes on genomic sequences of various organisms

研究代表者

宮崎　智（Miyazaki, Satoru）

東京理科大学・薬学部・教授

研究者番号：３０２９０８９４

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：ヒトやマウスの全ゲノム配列データのタンパク質遺伝子間領域やnon-coding RNA遺伝子の出現位置情報を抽出したデータベースを作成し、タンパク質遺伝子の発現制御領域配列中のシスエレメントの稼働率は30%程度であること、イントロン内在型non-coding遺伝子をもつタンパク質は、ヒトの脳や神経発達と関連性が高いこと、non-coding遺伝子とタンパク質遺伝子の発現制御制御領域は異なる1次配列構造を持つことを見出した。

研究成果の概要（英文）：The integration genomic databases available on INTRENET, we reconstructed our original database in which include the location of non-coding genes and control regions of the gene expression on genomic sequence, annotations and DNA sequence data. Bioinformatics approach with our original database allowed us to analyze some hidden rules on primary structure of genome sequences. In our approach suggested the following results; (1) Only 30% of cis-regulatory elements-like sequence in upper regions of coding genes are positive ones, (2)Proteins having intronic non-coding genes are high possibilities regarding with functions of the brain and neurons, (3)Primary structure of control regions of genome sequences are different between non-coding gene’s and coding gene’s.

研究分野：　応用ゲノム科学

キーワード：　ノンコーディングRNA　イントロン内ノンコーディングRNA

１．研究開始当初の背景

種々の完全長ゲノム配列が決定され、「ポストゲノム」といわれる時代となったが、全ゲノム配列に隠されている情報の解読は、まだ、第１ステージであるといっても過言ではない。ゲノム解析の多くは、タンパク質コーディング遺伝子の同定やその遺伝子が発現するタンパク質自身の機能解析が中心である。近年になってようやく、コーディング遺伝子の発現制御に関わるゲノム情報の解析が始められてきた。マイクロアレイ技術との協調もあり、ある疾病あるいは環境で協調的に働くと思われる遺伝子・タンパク質群の推定精度と規模が向上している。マイクロアレイ実験により示唆されたタンパク質群の関連性を説明するために、「共通の転写因子」に着目するのは、自然的な発想である。また、発現を制御する転写因子群の網羅的な解析が極めて重要であろうという見解は、iPS 細胞構築技術において導入された遺伝子の発見でも実証されてきている。ゲノム情報の中で、遺伝子の発現制御に関わる配列情報の解析は、ポストゲノム時代においてその重要度がますます増してきていると考えることができる。

遺伝子というと、タンパク質をコードしている coding 遺伝子に偏りがちであるが、一方では、non-coding 遺伝子（タンパク質以外の生体分子の遺伝子）の存在の役割が注目を集めている。その実像が十分に理解されているとはいえないが、small RNA を模倣して発現制御やタンパク質分解を制御する RNAi 技術が実用化されてきている実情もある。

RNA 関連遺伝子を含む non-coding 遺伝子の発現制御情報を含む統合データベースができれば、第 3 の医薬品として注目されている「核酸医薬品」の開発にも貢献できる。また、non-coding 遺伝子とその制御領域の進化についての研究も興味深い。

２．研究の目的

本研究では、small RNA を含む non-coding 遺伝子の発現制御に関わるゲノム配列パターンをバイオインフォマティクス手法で明らかにし、ゲノム配列の進化に伴う相違を検討することを目的とする。そのために、達成すべき小項目として、以下のようなもの想定した。

（１）国際塩基配列データベースや Ensembl データベースなど、公開データを駆使した non-coding 遺伝子の上流と下流の遺伝子間領域の網羅的な取得と更新の自動化。

（２）これまでの研究から、生化学実験によるシスエレメント様配列の長さは 20 塩基程度であるものの、エックス線立体構造データに登録されている、DNA とタンパク質の複合構造を検証したところ、実際の相互作用領域は、およそ 4 塩基から８塩基までの長さの塩基配列であるという知見を得ている。これをもとに、理論的に考えられる全てのパターン（４塩基長なら 4×4×4=256 種類）の配列の配列間領域での出現位置を網羅的に解析し、転写因子が相互作用を起こす可能性のある配列パターンをすべて観察する。また、各塩基配列パターンについて頻度分布などの統計情報の計算を行う。

（３）これまでに、いくつかの配列パターンの頻度分布を解析してみたところ、一様分布にはならず、出現頻度に２つの変極点をもつ分布が観測されているなど、文字列頻度になんらかの偏りが生じていることは明らかである。上記の理論パターンの中で、シスエレメントとして登録のある（coding 遺伝子のものも含む）パターンの出現頻度の統計的有意性の検討をおこなう。

（４）coding 遺伝子制御に関わる転写因子と non-coding 遺伝子の転写制御に関わる転写因子の比較を行う。

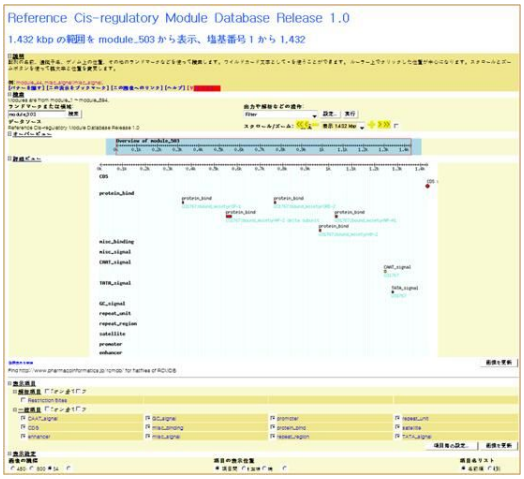
（５）ヒト、マウス、ラット、ショウジョウバエ、酵母での non-coding 遺伝子の上流・下流遺伝子間領域について、対応する配列領域の特定を行い、分子進化的手法をもちいて比較し、non-coding 遺伝子の発現制御や進化のメカニズムについての仮説を立案する。

以上の結果より、先行している coding 遺伝子の発現制御に関わるシスエレメント群と比較することで、non-coding 遺伝子発現制御の進化的メカニズム解析に言及することを最終目標とした。

３．研究の方法

（１）国際塩基配列データベースや Ensembl データベースなど、公開データを駆使した non-coding 遺伝子の上流と下流の遺伝子間領域の網羅的な取得と更新の自動化

研究代表者の宮崎は、これまでの研究において、DDBJ/EMBL/Genbank 国際塩基配列データベース中で、各々のタンパク質について生化学的な実験によって保証されているシスエメント配列を網羅的に取得してまとめ、そのタンパク質のゲノム上の位置を特定し、シスエレントのゲノムマップを作製してきている（下図：Reference cis-module database）。



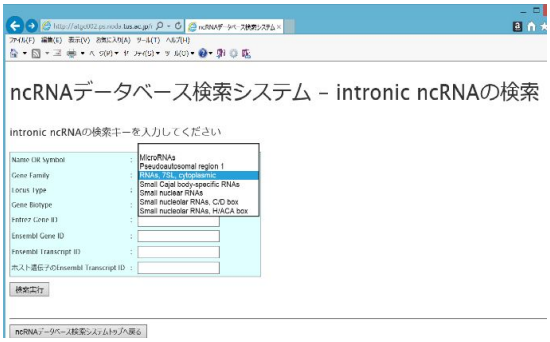

```
CIS_MODULE cis_module_429
DATE      31-JAN-2009
DEFINITION  Rat osteocalcin gene, complete cds.
SOURCE    M23637
FEATURES          Location/Qualifiers
   source                 /organism=Rattus norvegicus
            /mol_type=genomic DNA
            /db_xref=taxon:10116
            /tissue_lib=Clontech
   PROTEIN_BIND   /location=66..79
            /bound_moiety=NF1
   PROTEIN_BIND   /location=298..304
            /bound_moiety=AP2
   PROTEIN_BIND   /location=complement(324..331)
            /bound_moiety=AP1
   ENHANCER       /location=492..500
   PROTEIN_BIND   /location=complement(669..676)
            /bound_moiety=AP2
   ENHANCER       /location=complement(673..681)
   MISC_BINDING   /location=941..957
            /bound_moiety=cAMP
   CAAT_SIGNAL    /location=1003..1007
   TATA_SIGNAL    /location=1064..1070
   EXON           /location=1095..1191
            /number=1
   CDS            /location=join(1128..1191,1340..1372,1516..1585,1785..1917)
```

http://www.pharmacoinformatics.jp/cis/
フラットファイル形式

本研究では、この技術を活かし、non-coding 遺伝子に関して同様のリファレンスデータベースを作成する。一方で、non-coding 遺伝子の遺伝子間領域を網羅的な取得とデータベース化を行い、この２種のデータの公開を目指す。

（２）４塩基から８塩基までの長さの塩基配列について、理論的に考えられる全てのパターン（４塩基長なら４×４×４＝６４種類）の配列と全ゲノム配列の配列間領域のアライメントを行い、文字列の出現位置および頻度分布を作成する。

（３）上記の理論パターンの中で、シスエレメントとして登録のある（coding 遺伝子のものも含む）パターンの出現頻度の統計的有意性の検討からなる２つの課題を検討する。４から８塩基長の配列だけで、全てのシスエレメントが網羅できているとは言えないかもしれないが、エックス線解析によるタンパク質の立体構造データベースを参照して解析したところ、DNA とタンパク質の相互作用に関わる塩基部分配列長はおよそ４塩基前後であることが分かっている。そこで、本研究では、可能性が高い４塩基から８塩基を理論的に解析することとした。



（４）coding 遺伝子制御に関わる転写因子と non-coding 遺伝子の転写制御に関わる転写因子を比較する。

（５）Non-coding 遺伝子として公共データベースに登録がある代表的な遺伝子配列を用い、その上流に位置するシスエレメント配列を調べ、タンパク質遺伝子のシスエレメント配列との比較を行う。

４．研究成果

（１）データベースの作成と公開

３．（１）で述べたように、国際塩基配列データベース、ＥｎｓｅｍｂｌおよびＥＮＣＯＤＥデータベースの３種を使い、ヒトとマウスの全ゲノム配列情報から、non-coding 遺伝子の位置情報とその上流領域の情報を抽出した。Coding 遺伝子についても、ヒトとマウスの遺伝子間領域を網羅的に抽出し、解析用のデータセットを構築できた。その中で、long non-coding 遺伝子であって、タンパク質遺伝子のイントロンに存在するものについて、検索用のＧＵＩを構築し公開することができた。

（下図 http://atgc002.ps.noda.tus.ac.jp/ncrna/にてアクセス可能）。

（２）遺伝子間領域の配列頻度パターンについての解析

３．（２）（３）の方法によって、タンパク質遺伝子間領域における塩基配列パターンの解析を行ったところ、以下のような興味深い成果を得ることができた。

①シスエレメント様配列の稼働率

遺伝子間領域において、生化学的実験により証明されているシスエレメント配列についての、出現位置を特定した。それらについて、ＥＮＣＯＤＥプロジェクトによって、転写因子との結合が観測されたものとそうでないものに分け、シスエレメント配列と同じ塩基列をもつ配列の中で、実際に転写因子に認識される配列の割合（以下、シスエレメント配列の稼働率と呼ぶ。）を、転写因子毎に、まとめた（下表）。その結果、シスエレメント配列の稼働率は、80 件の転写因子でおおよそ 30%前後であることが判った。また、色を付けた転写因子のシスエレメント配列は、遺伝子間領域の中の反復配列中に多く存在することも判った。

転写因子名	反復配列	
	有	無
Myc	6.19	17.50
Nrsf	4.35	45.65
Pu1	9.48	9.25
Ap2a	2.81	5.12
Ap2c	0.59	2.51
Eraa	6.52	13.04
Erra	0.28	0.88
Fos	20.06	28.58
Fosl1	1.17	2.40
Jun	3.80	7.96
Junb	2.20	3.64
Tr4	2.14	23.38
Atf1	3.22	16.31
Atf2	1.11	5.71
Atf3	4.35	28.55
Bach1	7.29	13.63
Brca1	0.26	1.28
Cdp	0.00	2.17
Cebpb	9.98	13.70
Cebpd	0.80	0.93
Creb1	4.23	25.60
Ctcf	47.61	63.98
E2f1	0.26	4.32
E2f4	6.11	17.94
E2f6	4.67	27.30
Ebf1	18.42	22.35

転写因子名	反復配列	
	有	無
Egr1	66.12	58.10
Elf1	3.16	6.66
Elk1	1.59	12.36
Elk4	0.84	9.42
Ets1	0.22	1.21
Fosl2	10.70	19.16
Foxa1	7.00	9.71
Foxp2	0.73	1.66
Gata1	1.00	1.79
Gata2	1.57	2.99
Gata3	3.57	6.54
Gr	0.53	1.35
Hnf4a	9.43	14.38
Hnf4g	4.73	9.33
Hsf1	0.05	0.13
Irf1	1.19	1.71
Jund	32.01	38.76
Maff	49.00	50.99
Mafk	23.98	28.67
Max	5.86	16.61
Maz	4.89	14.53
Mef2a	0.08	0.89
Mef2c	0.48	2.91
Nanog	0.12	0.30
Nfe2	11.09	18.48
Nfic	1.81	5.79
Nfkb	7.21	15.25

転写因子名	反復配列	
	有	無
Nfya	23.27	1.45
Nfyb	50.10	21.29
Nrf1	25.79	77.92
P300	3.66	10.81
Pax5	1.52	8.21
Pou2f2	2.29	4.23
Pou5f1	4.32	5.56
Prdm1	2.88	3.87
Rfx5	0.58	1.36
Sp1	14.78	23.24
Sp2	5.20	11.28
Srebp1	0.04	0.27
Srebp2	0.13	0.14
Srf	24.43	19.61
Stat1	0.71	2.70
Stat3	2.86	4.99
Stat5a	0.73	1.95
Tal1	0.69	1.38
Tbp	0.58	1.60
Tcf7l2	4.88	11.25
Thap1	0.03	0.20
Usf1	15.03	17.80
Usf2	1.03	5.43
Yy1	1.53	2.35
Zbtb33	36.03	65.45
Zeb1	0.51	3.81
Znf263	42.38	21.04

②ヒトの 19113 個のタンパク質遺伝子

これらの遺伝子の上流に存在する稼働シスエレメントの位置の度数分布を基にしてクラスタリングを行ったところ、シスエレメントの存在パターンに対応して、７つのクラスタに分類された。この７つのクラスタに属するタンパク質遺伝子のアノテーション情報をＧＯ解析したところ、シスエレメントパターンは「生物学的プロセス」と相関している可能性が示唆された。

（３）intronic long non-coding RNA 遺伝子に係る成果について

本研究では、non-coding 遺伝子全般に係る特徴付の解析を目指してきたが、その中で、long non coding RNA であって、その遺伝子がタンパク質のイントロンに存在するものについて解析した成果について述べる。イントロンに存在する non-coding ＲＮＡをヒト、マウス、ラット、ハエ、線虫、酵母について調べ比較したところ、non-coding RNA の種類が、高等生物になるほど増えていく傾向があることが示唆された。miRNA のようにタンパク質の発現制御を司るものが、イントロンに存在していることが判明した。

ヒトとマウスのＸ染色体に存在するイントロン内在型 miRNA の配列情報から、miRNA の進化系統樹を作成したところ、ヒトおよびマウスのみの miRNA で１つのクラスタを形成しているノードがあった。このことから、ヒトとマウスの種分化に応じて、miRNA の機能分化が起こった可能性が示唆された。

さらに、ヒトとマウスにおいて、イントロン内在型 non-codingＲＮＡをもつ、タンパク質のアノテーションを用いて、ＧＯ解析をおこなったところ、ホスト遺伝子群について、生物学的プロセスからは、cell development（細胞発生）、neurogenesis（ニューロン新生）、nervous system development（神経系発達）、neuron projection development（神経投射発達）などと関わりがあることが示唆された。

また、細胞の構成要素からは、cytoskeleton（細胞骨格）、cell-cell junction（細胞間結合）などと関わりがあることが示唆された。さらに、MeSH 解析より、解剖カテゴリーからは Brain（脳）が、疾患カテゴリーからは telangiectasia（毛細血管拡張症）、Hereditary hemorrhagic（遺伝性出血性末梢血管拡張症）、Angelman Syndrome（Angelman 症候群：重度の精神遅滞・てんかんなどを主徴とする奇形症候群）などと関わりがあることが示唆された。細胞骨格は情報の伝達に重要な役割を果たしており、細胞骨格の破壊により脳の機能が障害される。その代表的な疾患にアルツハイマー病があり、それ以外にも細胞骨格が動脈硬化症、がん、精神疾患という異なる疾患に関与していることが報告されている。このことから、intronic ncRNA による遺伝子発現調節異常がこれらの疾患に繋がっている可能性がある。ヒトで顕著であった神経系・脳・細胞骨格などはいずれも非常に重要な役割を果たしているが、これらの特徴はマウスでは見られなかったことから、ncRNA による複雑な制御が生物の多様性に関与していることを示している。

（４）non-coding RNA の発現制御に係る研究の成果について

プロモーター領域に含まれる ATA-box、GC-box、CCAAT-box の数を比べると、miRNA、protein-coding gene ともに、TATA-box、CCAAT-box、GC-box の順で多く存在していた。また TATA-box の割合はどちらも高い結果となったが、GC-box と CCAAT-box の割合は protein-coding gene の方が miRNA より多い結果となった。miRNA gene と protein-coding gene のプロモーター領域の分布については、遺伝子上流の特定の位置に存在するようなプロモーターが確認できなかったため、規則性の解明には至らなかった。ヒト mirtron 上流 2kB では AAA と TTT で大きなピークが見られ、この配列パターンが多く用いられていることが分かる。

逆に AGC,ACG,TGC,TCG,GCG,CGC,CCG の位置はカウント数が極端に少なく、この種の塩基配列パターンがあまり用いられない。

また GGC,GCA,GCT,GCG,GCC の領域、CGA,CGT,CGG,CGC の領域は多くのグループでカウント数が小さい。この傾向はヒト miRNA

上流 2kB でも見られ、両者は非常によく似たパターンを示していた。マウス mirtron 上流 2kB においても、AAA、TTT に大きなピークが見られた。大まかな特徴はヒト mirtoron、miRNA の結果と同じである。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　5 件）

① “Protein functional site prediction using a conservative grade and a proximate grade.” J Data Mining Genomics Proteomics 6:175. (2015), Kondo Y, Miyazaki S doi:10.4174/2153-0602.1000175（査読有）
② “Genome-wide analysis of long noncoding RNA turnover.” Tani H1, Imamachi N, Mizutani R, Imamura K, Kwon Y, Miyazaki S, Maekawa S, Suzuki Y, Akimitsu N. Methods Mol Biol. 2014;1262:305-20. doi: 10.1007/978-1-4939-2253-6_19.(査読有)
③ “MeSH ORA framework: R/Bioconductorpackages to support MeSH over-representation analysis”, BMC BIOINFORMATICS, 2014, Koki Tsuyuzaki, Gota Morota, Manabu Ishii, Takeru Nakazato, Satoru Miyazaki and Itoshi Nikaido, doi:10.1186/s12859-015-0453-z, (査読有)
④ “Implication of bidirectional promoters containing duplicated GGAA motifs of mitochondrial function-associated genes”, AIMS Molecular Science,1, 1-26, 2013, Fumiaki Uchiumi, Makoto Fujikawa, Satoru Miyazaki and Sei-ichi Tanuma, doi:10.3934/molsci.2013.10. (査読有)
⑤ 「情報科学的手法による non-coding RNA の遺伝的発現制御の解析」, 情報処理学会研究報告バイオ情報学,2013-BIO-34,2,1-6, 2013,6,林 知里, 權 娟大, 宮崎 智 (査読無)

〔学会発表〕（計　8 件）

① 「文献注釈情報 MeSH を利用した網羅的な遺伝子の機能アノテーションパッケージ」、第 38 回 日本分子生物学会年会・第 88 回 日本生化学大会 合同大会、神戸ポートアイランド（兵庫県）、露崎弘慤、宮崎智、2015.12.1-4
② 「ヒトとマウスにおける転写制御情報を用いた p38qNARK 基質のクラスター分析」、第 38 回 日本分子生物学会年会・第 88 回 日本生化学大会 合同大会、神戸ポートアイランド（兵庫県）、阿子島圭、宮崎智、2015.12.1-4
③ “Functional site prediction of translation elongation factor 1A”, QBIC Workshop2015, Noda Campus, Tokyo Univ. of Science, Yosuke Kondo and Satoru Miyazaki, 2015.10.19
④ 「保存度と近接度を用いたタンパク質機能部位予測」、情報処理学会　研究報告バイオ情報学（BIO）　沖縄科学技術大学院大学、近藤洋介、權娟大、宮崎智、2015.6.23-25
⑤ 「複数の構造を用いた多重配列アライメント上のサイトの評価」近藤洋介、權娟大、深井文雄、宮崎智、生命医薬情報学連合大会、宮城県仙台市仙台国際センター、2014.10.2-4
⑥ 「共進化解析を利用した遺伝子間相互作用の予測」西川大貴、權娟大、宮崎智、第 37 回日本分子生物学会年会、パシフィコ横浜（神奈川県）、2014.11.25-27
⑦ 「micro-RNA gene と protein-coding gene のプロモーター領域の比較」, 今井はる奈, 權娟大, 宮崎智、日本薬学会第 134 年会 、熊本会場、2014.3.28-30
⑧ 「バイオインフォマティクス手法によるシスエレメント配列一次構造の網羅的解析」加納知佳、權娟大、宮崎智、日本薬学会第 134 年会、熊本会場、2014.3.28-30

〔その他〕
一般公開データベース
http://atgc002.ps.noda.tus.ac.jp/ncrna/

６．研究組織
(1)研究代表者
宮崎　智　（MIYAZAKI, Satoru）
東京理科大学・薬学部生命創薬科学科・教授
研究者番号：30290894