論文の内容の要旨

Discovery of microRNA genes and pseudogenes from genomic sequences

（ゲノム配列からのマイクロ RNA 遺伝子と偽遺伝子の発見）

寺　井　悟　朗

ゲノム配列には、遺伝子、制御領域、転移因子などの様々なタイプの配列要素が含まれている。ゲノムを詳細に理解するためには、それらの配列要素を可能な限り網羅的に発見することが重要である。生物学的実験はそれら配列要素を正確に発見することができる方法だが、多大な労力と費用を要する。そのため、ゲノム全体を生物学的実験で調べることは困難である。計算機的手法は、必ずしも信頼性が高いわけではないが、一般に高速であり、ゲノム全体を対象とした解析ができるという利点がある。したがって、計算機的手法は今やゲノムの解析には欠くことができないものとなっている。

本論文では、特定の配列要素を予測するための計算機的手法を 2 つ提案する。1 つ目はマイクロ RNA 遺伝子を予測する手法 miRRim2 である。マイクロ RNA はタンパク質をコードしない遺伝子の一種であり、ヒトにおいて数千個ものタンパクコード遺伝子の発現を制御していると言われている。また、マイクロ RNA は、癌を含むいくつかの疾病に関与していることが示されている。したがって、マイクロ RNA の発見は、生物学的にも医学的にも重要である。転写されたマイクロ RNA は特徴的な 2 次構造をとり、その 2 次構造は多くの場合で進化的に保存されている（図 1）。したがって、マイクロ RNA を正確に予測するためには、2 次構造と進化的特徴の両方を考慮することが重要である。我々の手法では、マイクロ RNA の各部位における 2 次構造と進化的特徴を多次元ベクトルで表現する。そして、多次元ベクトルの配列で表現されたマイクロ RNA を確率モデルによりモデル化する。miRRim2 はヒトゲノム全体を用いたクロスヴァリデーションテストにおいて、既存の手法による予測より優れた予測精度を達成した。さらに成熟マイクロ RNA の 5'末端の塩基位置を、感度・陽性的中率ともに 0.4 以上で予測することができた。

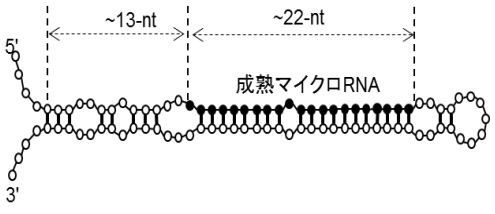

図 1　マイクロ RNA の模式的な 2 次構造
黒丸で示した塩基は、成熟マイクロ RNA となる部分である。この部分は他の部分よりも塩基対を多く形成し、かつ進化的に強く保存されている傾向がある。

2 つ目は、偽遺伝子を予測する手法 TSDscan である。偽遺伝子は転写された遺伝子のコピーであり、ほとんどの場合機能を持たないと考えられている。偽遺伝子は哺乳類のゲノムには非常に多く存在するため、その発見はゲノム配列のより正確な注釈付けに寄与する。これに加え、偽遺伝子の発見は親遺伝子の発現パターンに関する有用な知見をもたらす。たとえば、偽遺伝子を持つ遺伝子は生殖細胞（あるいは生殖細胞へと分化する細胞）で発現していると推測することができる。なぜなら、それらの細胞で生じた偽遺伝子だけが子孫に伝えられるからである。哺乳類の偽遺伝子周辺には 3 種類の配列特徴的がある（図 2)。TSDscan はこれらの配列特徴を利用して偽遺伝子を予測する。このアプローチはエクソン数やフレーム内終始コドンなどの偽遺伝子特徴を利用しないという点で、従来法とは明らかに異なっている。したがって、TSDscan は新しいタイプの偽遺伝子を発見できる可能性を持っている。実際に、TSDscan を用いてヒトゲノムを解析したことろ、300-bp 以下の短い偽遺伝子が多数存在することを発見した。さらに、発見した偽遺伝子と親遺伝子の長さを調査したところ、親遺伝子が長いほど短い偽遺伝子が生成されやすいという傾向を発見した。この観測結果から、我々は偽遺伝子の生成過程に関して次のような仮説を提案した：長い親遺伝子の転写物は、ほとんどの場合、ゲノムにコピーされる前に切断されている。切断された転写物はコピーの途中で速やかに分解されるため、短い偽遺伝子が生成される。

総括すると、我々は特定の配列要素を予測する計算機的手法を 2 つ開発した。1 つ目の手法 miRRim2 は、マイクロ RNA 遺伝子を高精度に予測できるだけでなく、その成熟体の位置も推定することができる。2 つ目の手法 TSDscan は、偽遺伝子周辺の配列特徴に基づき偽遺伝子を予測する。TSDscan を用いて、我々はヒトゲノムには短い偽遺伝子が多数存在することを明らかにした。これらの予測結果は、これら 2 種類の配列要素に関するより正確で網羅的な知見を与えるものであり、ゲノム配列のより詳細な理解に貢献する。

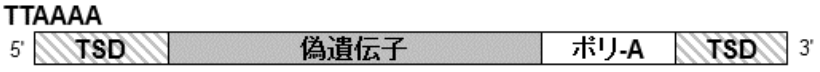

図２　偽遺伝子周辺の 3 つの配列特徴
偽遺伝子の両側にはターゲット部位重複（TSD）と呼ばれる短いタンデムリピートが存在する。また、3'末端にはポリ A が存在する。偽遺伝子の上流側には TTAAAA という配列があり、この配列は上流側 TSD と重なっている。

出典：

図１、２は、それぞれ文献［1］と［2］で出版された図を改変したものである。

［1］Terai G, Okida H, Asai K, Mituyama T. (2012) Prediction of conserved precursors of miRNAs and their mature forms by integrating position-specific structural features. *PLoS One* **7**, e44314.

［2］Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T. (2010) Discovery of short pseudogenes derived from messenger RNAs. *Nucleic Acids Res.* **38**, 1163-71