Retos en la predicción de genes de ARN no codificante

Hugo Armando Guillén Ramírez, Israel Marck Martínez-Pérez
Departamento de Ciencias de la Computación
Centro de Investigación Científica y de Educación Superior de Ensenada
Ensenada, B.C., México 22860
Email: hguillen@cicese.edu.mx, israelmp@cicese.mx

***Resumen*—El ADN se transcribe en varios tipos de ARN, pero no todo el ARN se traduce en proteínas. Este tipo de ARN se conoce como ARN no codificante (ARNnc). Si bien el estudio del ARNnc se ha dejado de lado en favor de la investigación de los genes que codifican proteínas, en la última década se convirtió en un área de investigación activa debido a las múltiples funciones del ARNnc. Actualmente, el descubrimiento de secuencias de ARNnc a un nivel genómico es una tarea computacionalmente costosa y con una baja precisión incluso en casos de prueba ideales. En el presente artículo se presenta una investigación en curso que pretende ayudar en el desarrollo o mejora de técnicas para el descubrimiento de este tipo de genes mediante búsqueda por homologías.**

I. Introducción

Anteriormente se creía que los productos génicos sólo eran proteínas, sin embargo también abarcan distintos tipos de ARN funcionales, también conocidos como ARN no codificantes, entre los que se encuentran los microARN, ARN de transferencia, ARN ribosomal y riboswitches. Con esto, se puede definir el término gen de ARN como una porción de ADN que codifica no para proteínas, sino para una cadena de ARN. Estas cadenas pueden presentar funciones catalíticas y regulatorias [1]. La existencia de los genes de ARN no codificante (ARNnc) fue propuesta simultáneamente junto a los genes codificantes de proteínas en 1961 por Jacob y Monod [2].

El problema de la predicción de genes es un problema desafiante y que todavía necesita mejoras, especialmente al analizar genomas grandes [3]. A pesar del avance estable del estado del arte, todavía no hay una manera de generar automáticamente modelos de genes de alta calidad para un genoma entero, incluso en uno tan estudiado como el humano [4]. Los métodos de predicción de ARNnc todavía están en su infancia comparados con los predictores de genes codificanes de proteínas, sin embargo, es un área en rápido crecimiento [5].

A. Predicción de genes

El problema de la predicción de genes consiste en identificar las porciones de ADN que son biológicamente funcionales. El problema puede enunciarse como uno de clasificación, donde el objetivo es etiquetar correctamente cada elemento de la secuencia de ADN como perteneciente a un exón, o una región no codificante (un intrón, región intergénica o región sin traducir). A continuación se presenta el enunciado formal del problema propuesto [6]:

- *Entrada:* una secuencia de ADN

$$X = (x_1, \ldots, x_n) \in \Sigma^*, \text{ donde } \Sigma = \{A, T, C, G\}. \tag{1}$$

- *Salida:* el etiquetado correcto de cada elemento en X perteneciendo a un exón, o a una región no codificante.

Por otro lado, el problema puede interpretarse como la extracción de exones putativos en una secuencia de ADN [7]:

- *Entrada:* una secuencia de ADN

$$X = (x_1, \ldots, x_n) \in \Sigma^*, \text{ donde } \Sigma = \{A, T, C, G\} \tag{2}$$

y una función de puntaje de exones $f(x)$.

- *Salida:* el conjunto de pares $(x_i, x_j), 1 \leq i < j \leq n$, tales que la subregión $X[i : j]$ representa un exón putativo y $f(X[i : j])$ es máximo.

Los métodos de predicción pueden descomponerse en dos partes: la recopilación de evidencia (detección de señales y sensado de contenido) y la integración de la evidencia mediante un modelo de estructura de genes [8].

B. Predicción de genes de ARNnc

Por lo general, los predictores de genes codificantes no toman en cosideración los intrones ni las regiones intergénicas. Sin embargo, estas regiones posiblemente contengan alguna función reguladora. Los métodos para identificar y clasificar el ARNnc en estas regiones utilizan diferentes estrategias basadas en las características soportadas por evidencia experimental e *in silico*. Estas estrategias se dividen en tres grandes grupos: *de novo*, búsqueda por homologías y búsqueda específica por familias.

1) De novo: Los predictores *de novo* se basan solamente en las características del genoma. Por esto, la estrategia general que siguen consiste en utilizar ventanas y predecir el plegamiento del ARN en la ventana utilizando un modelo de energía libre [9]. Sin embargo, se ha demostrado que las estructuras predichas de los genes de ARN no son significantemente más estables que secuencias aleatorias generadas a partir del trasfondo genómico, por lo que no puede usarse ese criterio para distinguirlos [10]. Como resultado, los esfuerzos se han enfocado en integrar al modelo información de distintas fuentes que puedan generar una estructura secundaria consenso [11].

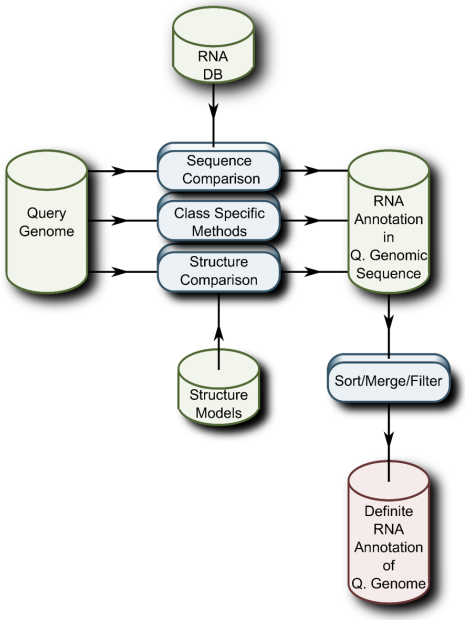

Fig. 1. *Pipeline* para la predicción de ARNnc [12].

2) Búsqueda por homologías y búsqueda específica de familias: La búsqueda por homologías consiste en utilizar familias de ARN conocidas en una porción de algún genoma. En el caso de la búsqueda específica de familias, los métodos se basan en entrenar modelos de aprendizaje máquina dedicados a ese tipo de ARN.

La búsqueda por homologías (al igual que la predicción de novo) está íntimamente relacionada con los problemas estructurales en el ARN: predicción del plegamiento y alineamientos estructurales. De forma bastante genérica, un predictor podría seleccionar una ventana de n-nucleótidos, tratar de predecir la estructura secundaria de esa secuencia de tamaño n, y por medio de alineamiento estructural evaluar la similitud con las familias que contenga la base de datos. Sin embargo, a pesar de que la estructura secundaria es importante como parte de la identificación, la estabilidad de muchas estructuras secundarias de ARNnc no es lo suficientemente diferente de la estabilidad predicha de una secuencia aleatoria para ser útil como un enfoque predictor de genes general [10]. Por ello, los métodos de búsqueda por homologías utilizan diferentes evidencias biológicas para generar las estructuras secundarias ya sea de las familias, o de la consulta. Entre estos métodos se encuentra la unión de la información procedente de alineamientos a nivel primario y la generalización de los modelos ocultos de Markov llamada gramáticas estocásticas libres de contexto (SCFG, por sus siglas en inglés).

En la práctica, es posible unir y complementar los enfoques *de novo*, específicos y por homologías para desarrollar un flujo de trabajo (*pipeline*) para la predicción [12] (Figura 1).

II. Trabajo previo

Debido a las particularidades de los genes ARNnc, los criterios comúnmente considerados al intentar detectarlos son los siguientes:

- Un gran número de clases conocidas de ARNnc no presentan marcos de lectura abiertos largos.
- Sus secuencias tienen codones de terminación inesperados.
- Las moléculas de ARN usualmente presentan conservación en su estructura secundaria y raramente en su estructura primaria.
- Algunos ARNnc conocidos tienen estructuras tridimensionales complejas, así como funciones catalíticas o estructurales.

Estas características impiden el uso de las homologías desarrolladas para la clasificación de genes codificantes de proteínas [13], por lo que un enfoque común en los predictores es identificar los genes de ARNnc utilizando estructuras secundarias y motivos. Ejemplos de estas herramientas incluyen tRNAscan-SE [14] y Snoscan [15], y la propuesta por [16] para el caso de microARN (miRNA).

Un enfoque más general para resolver el problema consiste en primero alinear las secuencias de nucleótidos (genómicas, RNA-seq y ESTs) de organismos relacionados cercanamente al genoma objetivo y entonces buscar señales de estructuras secundarias conservadas. No obstante, este es un proceso complejo que todavía arroja una alta tasa de falsos positivos y requiere recursos computacionales sustanciales [5]. Ejemplos de estas herramientas son qRNA [17], StemLoc [18], pmcomp y pmmulti [19] e Infernal el cual utiliza la base de datos Rfam [20]. Infernal [21] es un predictor que identifica ARNnc mediante la búsqueda de las estructuras secundarias de ARN. Este predictor construye perfiles de consensos entre las estructuras espaciales de ARN, conocidos como modelos de covarianza, con el fin de encontrar similitudes entre la secuencia investigada y la estructura de consenso secundaria de cada una de las familias de ARN almacenadas en la base de datos Rfam [20]. A pesar de ser el mejor predictor de ARNnc multiclase, en un caso de prueba de 10.16MB alcanza una precisión menor al 85% en 4359 horas de CPU [22].

La herramienta tRNAscan [14] se considera uno de los predictores de ARN de transferencia (ARNt) más preciso. Su método consiste en combinar tres programas: dos predictores de ARNt y el modelo de covarianza de [21] entrenado previamente con secuencias de ARNt. La ejecución de los tres programas resulta en un identificador de ARNt con una alta sensibilidad (99% - 100%) y alta especificidad al presentar una tasa de falsos positivos menor a 7.0e-5 por Mb en una velocidad razonable. Portrait [23] identifica ARNnc mediante una máquina de soporte de vectores en transcriptomas incompletos o de especies no totalmente caraterizadas. El resultado de este programa es la probabilidad de que un transcrito sea o no un codificador de proteínas.

Vienna [24] es un conjunto de paquetes usados para generar o comparar estructuras secundarias de ADN. El plegado en esta herramienta usa algoritmos de predicción basados en la energía libre del ARN y la probabilidad del pareado de bases [25]. De manera particular, el paquete RNAfold [24], [26] se basa en la hipótesis de que la molécula de ARN se pliega en la estructura termodinámica más estable, la que tiene mínima energía libre. RNAmmer [27] predice ARN ribosomal (ARNr) utilizando la unidad ribosomal 5S y la base de datos de ARNr Europea para

generar alineamientos estructurales, los cuales se utilizan para la construcción de librerías de cadenas de Markov.

trCYK [28] y RNA-CODE [29] son propuestas recientes que se enfocan en predecir ARNnc utilizando como entrada las lecturas generadas por los secuenciadores de nueva generación. Finalmente, ncRNA-Agents [30] es un predictor que utiliza predictores existentes mediante un enfoque de multiagentes.

Las bases de datos de ARNnc son una parte fundamental de la predicción. Estas bases de datos son creadas a partir de datos experimentales y computacionales, y las que se utilizan comúnmente son NONCODE [31], RNAdb [32], miRBase [33], snoRNA [34], fRNAdb [35] y Rfam [20].

III. Retos en la predicción de genes de ARNnc

A continuación se presentan algunos de los retos presentes en el problema de la predicción de genes de ARNnc.

A. *Conservación*

Una secuencia de ARN conservada es aquella que es altamente similar entre varias especies. Un ejemplo de ARN altamente conservado son los miRNA y los snoRNA [13]. Sin embargo, la gran mayoría de los genes de ARN no son conservados a nivel primario, sino a nivel secundario y terciario [36]. Incluso cuando las secuencias están conservadas a nivel primario, las homologías de nucleótidos no son tan fáciles de detectar como las homologías en proteínas, lo cual limita el poder de los enfoques basados en evidencia y hace necesario el uso de modelos de clasificación [29]. Es por esto que para la predicción de ARNnc, los biólogos utilizan diferentes herramientas bioinformáticas con distintos métodos, para después hacer un análisis combinado de todas las evidencias recabadas y decidir si la secuencia es un ARNnc potencial [30].

B. *Carencias en la anotación*

Los segmentos de ARN no codificantes largos (lncRNA) no están bien anotados, debido a que faltan las posiciones de inicio y término; medir la similitud a nivel primario no es suficiente debido a la falta de conservación de los genes ARNnc; y la selección de un patrón de búsqueda de una familia en específico requiere de un experto que lo evalúe manualmente [37].

No existe un genoma completamente anotado en genes de ARNnc, por lo que no es posible estimar del todo la precisión y sensibilidad del predictor; muchos ARNnc son de longitud pequeña (< 100nt), sin embargo en general son de longitud variable; y a la hora de entrenar un modelo de reconocimiento probabilístico, si se excluye una secuencia cuya mutación es realmente una característica de la familia, esta no será considerada en la búsqueda [36]. Además, es difícil encontrar homólogos fuera del rango filogenético de ejemplos conocidos [38].

C. *Dificultad computacional*

Actualmente, la búsqueda y la predicción de estructuras tienen un alto costo computacional, además, la mayoría de los métodos no consideran pseudonudos a pesar de que existen y

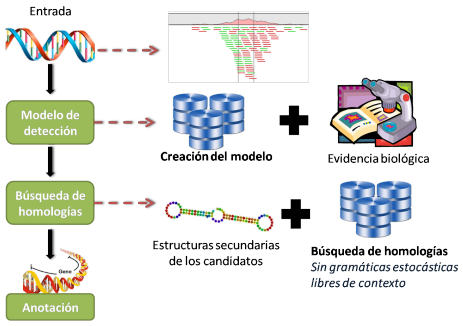

Fig. 2. Esquema de la propuesta de trabajo.

tienen funciones dentro de las familias de ARNnc; y el método estándar de comparar predicciones hacia algún *estándar de oro* es problemático debido a que de los verdaderos positivos disponibles, esto es, ARNnc conocidos, a pesar de que son bien recuperados por los métodos, pueden no ser representativos de ARNnc sin descubrir [11]. Existen ciertas pistas biológicas que pueden ayudar a los predictores: el uso de codones, sustituciones sinónimas y no sinónimas, y la energía de mínimo plegamiento [25]. Incluso con esta información sigue siendo difícil distinguir entre los genes ARNnc reales, falsos y genes codificadores de proteínas pobremente conservados que producen pequeños péptidos, especialmente en los casos de regiones intergénicas largas de ARNnc (lincRNA) [39]–[41] y pseudogenes expresados [42].

IV. Propuesta de trabajo

El trabajo de investigación propuesto consiste en desarrollar un predictor de genes ARNnc por búsqueda de homologías en familias conocidas de ARNnc, apoyado por clasificación, metaheurísticas, la evidencia biológica del estado del arte y las características intrínsecas de las secuencias de ARN. La búsqueda se llevará a cabo evitando el uso de las gramáticas estocásticas libres de contexto debido a su alto costo computacional.

El esquema de la solución se muestra en la Figura 2. La primer fase de la investigación apunta hacia la creación de un modelo que diferencíe algunas familias de ARNnc mediante su información estructural. La segunda fase es la mejora de los algoritmos de búsqueda en la base de datos del modelo utilizando heurísticas para acelerar el procesamiento. Respecto a la evaluación del método se utilizarán los casos de prueba para Infernal.

V. Conclusiones

Se presentó el problema de la predicción de genes codificantes y genes de ARNnc, así como los retos más importantes en la solución de este último. La predicción de genes de ARNnc muestra ser un problema desafiante, por lo que se presenta clara la oportunidad de mejorar los métodos actuales tanto en tiempo de ejecución como en precisión de resultados.

Agradecimientos

Los autores agradecen a los revisores anónimos por sus comentarios constructivos y sus valiosas sugerencias, así como al Consejo Nacional de Ciencia y Tecnología por el apoyo económico brindado durante el desarrollo de la investigación.

REFERENCIAS

[1] A. Serganov y D. J. Patel, "Ribozymes, riboswitches and beyond: regulation of gene expression without proteins," *Nature Reviews Genetics*, vol. 8, no. 10, pp. 776–790, 2007.

[2] F. Jacob y J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of molecular biology*, vol. 3, no. 3, pp. 318–356, 1961.

[3] S. Maji y D. Garg, "Progress in gene prediction: Principles and challenges," *Current Bioinformatics*, vol. 8, no. 2, pp. 226–243, 2013.

[4] T. Alioto, "Gene prediction," en *Evolutionary Genomics*. Springer, 2012, pp. 175–201.

[5] M. Yandell y D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nature Reviews Genetics*, vol. 13, no. 5, pp. 329–342, 2012.

[6] S. Bandyopadhyay, U. Maulik, y D. Roy, "Gene identification: classical and computational intelligence approaches," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 38, no. 1, pp. 55–68, 2008.

[7] W. H. Majoros, *Methods for computational gene prediction*. Cambridge University Press Cambridge, 2007, vol. 1.

[8] C. Mathé, M.-F. Sagot, T. Schiex, y P. Rouzé, "Current methods of gene prediction, their strengths and weaknesses," *Nucleic acids research*, vol. 30, no. 19, pp. 4103–4117, 2002.

[9] J. Gorodkin y I. L. Hofacker, "From structure prediction to genomic screens for novel non-coding rnas," *PLoS computational biology*, vol. 7, no. 8, p. e1002100, 2011.

[10] E. Rivas y S. R. Eddy, "Secondary structure alone is generally not statistically significant for the detection of noncoding rnas," *Bioinformatics*, vol. 16, no. 7, pp. 583–605, 2000.

[11] J. Gorodkin, I. L. Hofacker, E. Torarinsson, Z. Yao, J. H. Havgaard, y W. L. Ruzzo, "De novo prediction of structured rnas from genomic sequences," *Trends in biotechnology*, vol. 28, no. 1, pp. 9–19, 2010.

[12] J. Gorodkin y I. L. Hofacker, "From structure prediction to genomic screens for novel non-coding rnas," *PLoS computational biology*, vol. 7, no. 8, p. e1002100, 2011.

[13] K. C. Pang, M. C. Frith, y J. S. Mattick, "Rapid evolution of noncoding rnas: lack of conservation does not mean lack of function," *Trends in Genetics*, vol. 22, no. 1, pp. 1–5, 2006.

[14] T. M. Lowe y S. R. Eddy, "trnascan-se: a program for improved detection of transfer rna genes in genomic sequence," *Nucleic acids research*, vol. 25, no. 5, pp. 0955–964, 1997.

[15] P. Schattner, A. N. Brooks, y T. M. Lowe, "The trnascan-se, snoscan and snogps web servers for the detection of trnas and snornas," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W686–W689, 2005.

[16] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge *et al.*, "Prediction of mammalian microrna targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[17] S. R. Eddy, "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an rna secondary structure," *BMC bioinformatics*, vol. 3, no. 1, p. 18, 2002.

[18] I. Holmes y G. Rubin, "Pairwise rna structure comparison with stochastic context-free grammars." en *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2002, p. 163.

[19] I. L. Hofacker, S. H. Bernhart, y P. F. Stadler, "Alignment of rna base pairing probability matrices," *Bioinformatics*, vol. 20, no. 14, pp. 2222–2227, 2004.

[20] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, y S. R. Eddy, "Rfam: an rna family database," *Nucleic acids research*, vol. 31, no. 1, pp. 439–441, 2003.

[21] S. R. Eddy y R. Durbin, "Rna sequence analysis using covariance models," *Nucleic acids research*, vol. 22, no. 11, pp. 2079–2088, 1994.

[22] E. P. Nawrocki y S. R. Eddy, "Infernal 1.1: 100-fold faster rna homology searches," *Bioinformatics*, vol. 29, no. 22, pp. 2933–2935, 2013.

[23] R. T. Arrial, R. C. Togawa, y M. M. Brigido, "Screening non-coding rnas in transcriptomes from neglected species using portrait: case study of the pathogenic fungus paracoccidioides brasiliensis," *BMC bioinformatics*, vol. 10, no. 1, p. 239, 2009.

[24] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, y P. Schuster, "Fast folding and comparison of rna secondary structures," *Monatshefte für Chemie/Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[25] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for rna secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

[26] I. L. Hofacker, M. Fekete, y P. F. Stadler, "Secondary structure prediction for aligned rna sequences," *Journal of molecular biology*, vol. 319, no. 5, pp. 1059–1066, 2002.

[27] K. Lagesen, P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes, y D. W. Ussery, "Rnammer: consistent and rapid annotation of ribosomal rna genes," *Nucleic acids research*, vol. 35, no. 9, pp. 3100–3108, 2007.

[28] D. L. Kolbe y S. R. Eddy, "Local rna structure alignment with incomplete sequence," *Bioinformatics*, vol. 25, no. 10, pp. 1236–1243, 2009.

[29] C. Yuan y Y. Sun, "Rna-code: A noncoding rna classification tool for short reads in ngs data lacking reference genomes," *PloS one*, vol. 8, no. 10, p. e77596, 2013.

[30] W. Arruda, C. G. Ralha, T. Raiol, M. M. Brígido, M. E. M. Walter, y P. F. Stadler, "ncrna-agents: A multiagent system for non-coding rna annotation," en *Advances in Bioinformatics and Computational Biology*. Springer, 2013, pp. 136–147.

[31] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao, y R. Chen, "Noncode: an integrated knowledge database of non-coding rnas," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D112–D115, 2005.

[32] K. C. Pang, S. Stephen, M. E. Dinger, P. G. Engström, B. Lenhard, y J. S. Mattick, "Rnadb 2.0: an expanded database of mammalian non-coding rnas," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D178–D182, 2007.

[33] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman, y A. J. Enright, "mirbase: microrna sequences, targets and gene nomenclature," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D140–D144, 2006.

[34] L. Lestrade y M. J. Weber, "snorna-lbme-db, a comprehensive database of human h/aca and c/d box snornas," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D158–D162, 2006.

[35] T. Mituyama, K. Yamada, E. Hattori, H. Okida, Y. Ono, G. Terai, A. Yoshizawa, T. Komori, y K. Asai, "The functional rna database 3.0: databases to support mining and annotation of functional rnas," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D89–D92, 2009.

[36] P. Menzel, J. Gorodkin, y P. F. Stadler, "The tedious task of finding homologous noncoding rna genes," *RNA*, vol. 15, no. 12, pp. 2075–2082, 2009.

[37] J. Gorodkin, I. L. Hofacker, y W. L. Ruzzo, "Concepts and introduction to rna bioinformatics," en *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*. Springer, 2014, pp. 1–31.

[38] E. K. Freyhult, J. P. Bollback, y P. P. Gardner, "Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding rna," *Genome research*, vol. 17, no. 1, pp. 117–125, 2007.

[39] T. R. Mercer, M. E. Dinger, y J. S. Mattick, "Long non-coding rnas: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[40] S. van Leeuwen y H. Mikkers, "Long non-coding rnas: guardians of development," *Differentiation*, vol. 80, no. 4, pp. 175–183, 2010.

[41] T. Hung y H. Y. Chang, "Long noncoding rna in genome regulation: prospects and mechanisms," *RNA biology*, vol. 7, no. 5, pp. 582–585, 2010.

[42] O. H. Tam, A. A. Aravin, P. Stein, A. Girard, E. P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R. M. Schultz *et al.*, "Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes," *Nature*, vol. 453, no. 7194, pp. 534–538, 2008.