Modelagem de Redes Regulatórias para a Descoberta de Novos Biomarcadores de Doenças Complexas

Rafael Pompeu[1,2], Leandro Magalhães[2,3], Ândrea Ribeiro-dos-Santos[2,3], Amanda Vidal[2,3], Gilderlanio S. Araújo[2,3]

[1]Faculdade de Computação e Telecomunicação – Universidade Federal do Pará (UFPA)
Belém – PA – Brasil

[2]Laboratório de Genética Humana e Médica – Universidade Federal do Pará (UFPA)
Belém – PA – Brasil

[3]Pós-Graduação em Genética e Biologia Molecular
Universidade Federal do Pará (UFPA) Belém – PA – Brasil

{rafaelpompeu988, gilderlanio}@gmail.com

Abstract. *This study presents a proposal for modeling regulatory networks based on public biological data from three classes of regulatory elements (ncRNAs) such as miRNAs, circRNAs and piRNAs and their relations (regulation and origin) with genes. The integrated networks and its association with clinically relevant biological data allowed the discovery of potential candidate genes related to complex diseases. According to this network's centrality and neighborhood overlap, we identified novel potential genes and ncRNAs that could act as biomarkers of complex diseases considered major public health issues.*

Resumo. *Este estudo apresenta uma proposta de modelagem de redes regulatórias extraindo dados biológicos públicos de três classes de elementos regulatórios (ncRNAs), tais como os miRNAs, circRNAs e piRNAs e suas relações (regulação e origem) com genes. A integração das redes de ncRNAs e sua associação com dados biológicos de importância clínica permitiu a identificação de genes candidatos que potencialmente atuam no desenvolvimento de doenças complexas. A partir das características de centralidade e sobreposição de ligações desta rede, identificamos novos genes e ncRNAs potenciais biomarcadores de doenças complexas consideradas graves problemas de saúde pública.*

1. Introdução

Os RNAs não-codificantes (ncRNAs) são uma classe de RNAs estruturais e regulatórios que não são traduzidos em proteínas, e que representam cerca de 70% do genoma humano [Morris and Mattick 2014] [Costa 2010]. Apesar de vários estudos funcionais sugerirem que a desregulação da expressão dos ncRNAs está diretamente associada com o desenvolvimento de doenças complexas (tais como o câncer e a diabetes do tipo 2), a função biológica da maior parte destes ainda permanece pouco conhecida [Esteller 2011].

Em termos numéricos, os RNAs circulares (circRNAs), piwi-interacting RNAs (piRNAs) e microRNAs (miRNAs) estão entre os tipos de ncRNAs regulatórios de maior representatividade em humanos [Bahn et al. 2015]. Supõe-se que tamanha abundância

reflita o pouco conhecido espectro de ação destes ncRNAs. As interações existentes entre os ncRNAs e seus genes de origem ou genes-alvo são extremamente complexas e dinâmicas, e faz-se necessária a implementação de métodos computacionais para permitir o seu melhor entendimento [Volinia et al. 2010].

A modelagem de redes complexas representa uma ferramenta importante para permitir a visualização e mensuração das interações entre diferentes componentes [Girvan and Newman 2002]. As redes bipartidas são uma classe que relaciona conjuntos de elementos distintos, nos quais os elementos de um mesmo conjunto não fazem ligações entre si — elas têm sido amplamente aplicadas na biologia para a representação de ligações enzima-reação em vias metabólicas, associações gene-doença e rede ecológicas, por exemplo [Kontou et al. 2016] [Burgos et al. 2008].

A aplicação de redes bipartidas no estudo dos ncRNAs pode elucidar a interação dos ncRNAs com os genes, bem como elucidar os mecanismos moleculares em que estão envolvidos. Além disso, as redes de interação são capazes de apontar os ncRNAs que podem ser eventualmente utilizados como biomarcadores de risco ao desenvolvimento de doenças [Shankaraiah et al. 2018].

A identificação de biomarcadores de risco ao desenvolvimento de doenças tem sido alvo de inúmeros estudos funcionais, uma vez que permite o diagnóstico precoce, a prevenção e o monitoramento da evolução do quadro clínico de pacientes [Maegdefessel 2014]. Portanto, as consequências imediatas da identificação de biomarcadores específicos são a diminuição da incidência de doenças e dos índices de mortalidade, o que gera um grande impacto positivo nos índices de saúde pública a nível mundial.

Diante do cenário proposto, o objetivo deste trabalho é construir redes de interação com todos os miRNAs, piRNAs e circRNAs descritos em humanos, tendo como foco os genes regulados pelos miRNAs e os genes que produzem os piRNAs e os circRNAs. A integração destas três redes de interação permitiu a identificação de genes humanos que estão relacionados à doenças complexas e que apresentam o maior número de ncRNAs relacionados. Estes genes e seus ncRNAs relacionados são potenciais biomarcadores de risco ao desenvolvimento de doenças complexas.

Além da presente introdução, este artigo está dividido nas seguintes seções: a seção 2 relata os trabalhos relacionados; a seção 3 descreve a modelagem da rede, bases de dados e ferramentas adotadas no desenvolvimento do trabalho; a seção 4 expõe os resultados encontrados, sendo os genes candidatos no processos de doencas e a seção 5 descreve as conclusões finais.

2. Trabalhos relacionados

Os ncRNAs são elementos regulatórios abundantes em humanos, mas ainda poucos explorados como potenciais candidatos a terapias gênicas [Rasool et al. 2016]. Dentre os ncRNAs, os miRNAs são os mais estudados e os únicos que possuem uma ferramenta para análise integrada com redes e métricas de redes complexas [Nair et al. 2014]. Devido à sua recente descoberta, os piRNAs e os circRNAs ainda são classes pouco estudadas e subrepresentadas em estudos funcionais [Vidal et al. 2016]. Entre as diversas bases de dados, apenas o miRNET (www.mirnet.ca) [Fan et al. 2018] destaca-se como uma ferramenta web que integra, além dos dados genômicos de miRNAs, as métricas de

centralidade, tais como o grau dos elementos, betweenness e o caminho geodésico dos nós (caminhos curtos). O miRNET é a única ferramenta que consegue relacionar interações funcionais de miRNAs com doenças e com outras classes de ncRNAs.

Como ponto inicial, este trabalho integra dados genômicos relacionados a três classes de ncRNAs, diferentemente dos trabalhos citados anteriormente que têm propósito de estudar cada classe separadamente. Neste trabalho, métricas de centralidades e sobreposição são integradas a dados clínicos com o intuito de identificar elementos que potencialmente tem um papel biológico de impacto. Ainda, este trabalho se diferencia por identificar pares de genes que compartilham o mesmo conjunto de ncRNAs, resultando em genes candidatos que contribuem no processo de doenças complexas.

3. Metodologia

Os sistemas biológicos envolvem diversos elementos genéticos e moleculares que desencadeiam doenças complexas, tais como câncer e doenças neuropsiquiátricas. Dessa forma, a metodologia aplicada para conduzir esta pesquisa baseia-se na integração de bases de dados públicos de ncRNAs e extração de relações com genes, agregadas à bases de dados clínicas. Com o intuito de priorizar elementos candidatos como alvos de importância clínica e terapêutica, aplicamos métricas de redes complexas como forma de identificar potenciais candidatos.

3.1. Bases de dados de elementos regulatórios

Entre as bases de dados públicas de elementos regulatórios, extraimos os arquivos de dados tabulados das três principais relativas a cada ncRNA: a) miRTarBase [Chou et al. 2017] é um banco de dados que armazena os genes-alvo regulados pelos miRNAs e como essas interações foram validadas experimentalmente; b) piRBase [Wang et al. 2018], que cataloga os genes de origem de todos os piRNAs descritos em humanos; e c) circBase [Glažar et al. 2014], que mapeia as regiões gênicas que produzem os circRNAs. A Tabela 1 sumariza a abundância de elementos regulatórios e número de genes encontrados em cada base de dados.

Após extração dos dados, encontramos 2599 miRNAs que "regulam" 15.064 genes, os quais foram filtrados para permanecer apenas aqueles que foram validados por mais de um experimento, restando 648 miRNAs que "regulam" 2.338 genes; aproximadamente 21 mil piRNAs que são "produzidos" por 4.295 mil genes e $\sim$91 mil circRNAs que são originados de 11.813 genes.

Tabela 1. Sumário do número de elementos regulatórios (ncRNAs), genes encontrados em cada base de dados e o número de interações entre os genes e os ncRNAs.

Database	ncRNA	#ncRNA	#Genes	Interactions	#Interactions
MiRTarBase	miRNA	648	2.338	"regulate"	6.708
piRBase	piRNA	7.948	4.295	"produce"	21.277
circBase	circRNA	91.032	11.813	"produce"	91.032

3.2. Base de dados clínicos

O NGHRI/EBI GWAS Catalog [Buniello et al. 2018] é a principal fonte de dados de estudos de associação de fatores genéticos com doenças e outros traços, em escala genômica. Durante a realização deste trabalho, a base de dados continha 3.764 estudos de associação e 107.785 associações distintas entre variantes genéticas e doenças complexas.

3.3. Modelagem da Redes

A rede gene-ncRNA (G) é um grafo bipartido $G = (V, U, E)$, no qual o conjunto de nós é composto pelos genes (V) e pelos ncRNAs (U) e (E) representa o conjunto das arestas. As interações extraídas das bases de dados (ver Tabela 1) geram dois tipos de arestas: a) "regula", para representar a interação de regulação gênica entre miRNAs e genes e b) "produz", para representar o mapeamento entre o gene e a origem genômica dos piRNAs e circRNAs (ver modelo da Figura 1).

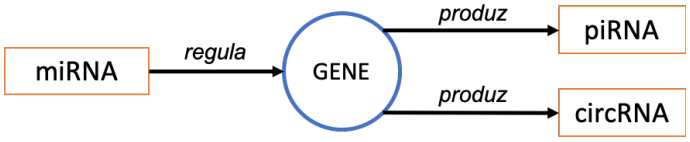

Figura 1. Representação gráfica da rede gene-ncRNAs.

3.4. Métricas de Redes Complexas

Para medir a centralidade dos elementos da rede e a sobreposição de elementos regulatórios, fizemos uso das seguintes métricas:

1. **Grau.** O grau do nó representa o número observado de ligações de um nó com outros nós.
2. **Betweenness.** Em um grafo, a métrica de centralidade betweenness para o vértice v é obtida a partir da proporção de caminhos curtos compartilhados, que passa pelo vértice E, definido pela equação (1), no qual $\mid G \mid$ é o números de elementos na rede, g_{uvk} é o caminho geodésico que passa por u, v e k, g_{uv} passa somente por u e v sendo B_{max} a normalização dada para cada conjunto V e U [Barthelemy 2018].

$$B_v = \frac{1}{2} \frac{\sum_{u \neq v}^{|G|} \sum_{k \neq u,v}^{|G|} \frac{g_{uvk}}{g_{uv}}}{B_{max}} \tag{1}$$

3. **Vizinhos comuns.** O número de vizinhos comuns é uma métrica simples e diz respeito a como dois nós se agrupam na rede, capturando o numero de elementos que são ligados ao par de elementos. Uma caracteristica que pretendemos investigar são os pares de genes que são regulados ou produz os mesmos elementos regulatórios. Estes pares de genes podem atuar no mesmo processo biológico. O número de vizinhos pode ser computado por uma operação de conjuntos definida na equação (2).

$$C_N = N(u) \cap N(v) \tag{2}$$

3.5. Enriquecimento Funcional

Para determinar as funções e vias biológicas dos genes mais importantes das redes geradas, realizou-se uma análise de enriquecimento funcional pela ferramenta web STRING v11.0 (https://string-db.org).

4. Resultados

Analisando o grau de conectividade das redes, notamos que as distribuições seguem a *power law* (Figura 2A), indicando que tanto os genes quanto os ncRNAs estão ligados a outros nós na rede por poucas interações. Esta característica da rede indica um discreto e pequeno conjunto de elementos centrais com muitas interações (*hubs*). Analisandos as distribuições de grau do ponto de vista biológico, atestamos que os genes são regulados por diferentes conjuntos de miRNAs e produz um número diverso de piRNAs e circRNAS.

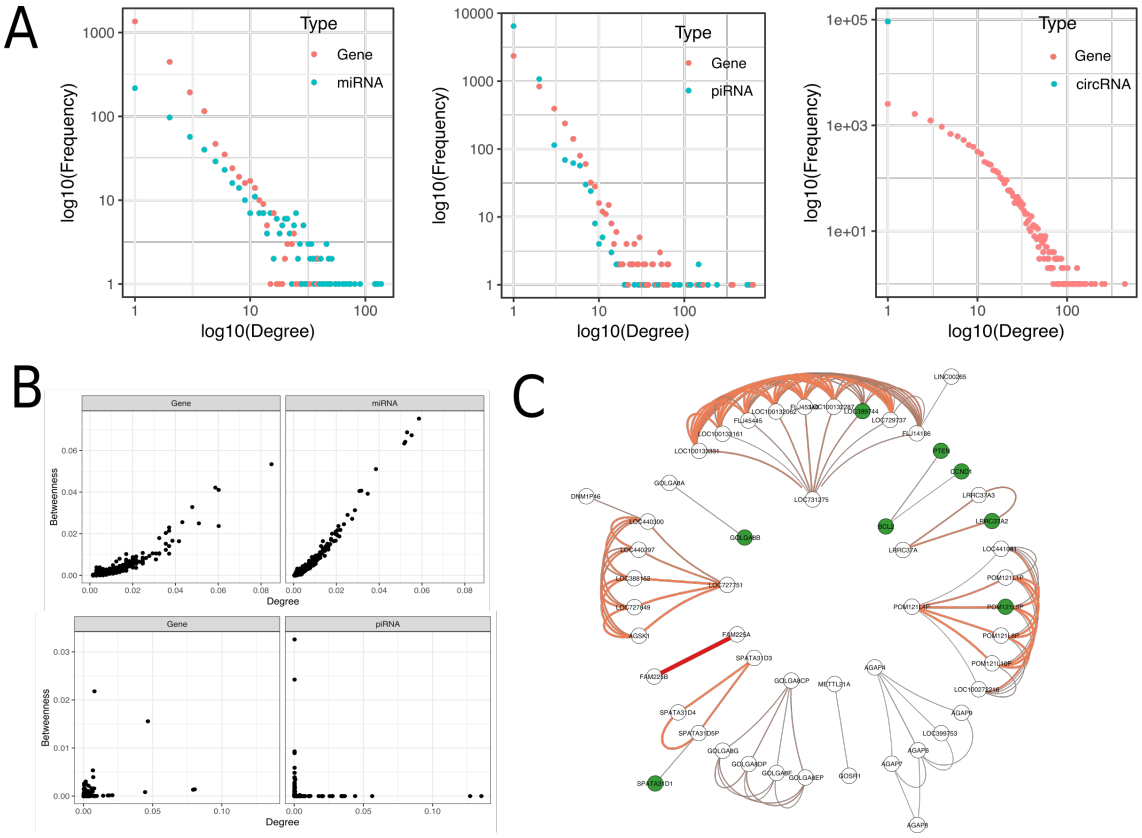

Figura 2. Distribuição do grau para cada elemento da rede considerando as diferentes classes de ncRNA. B) Relação entre o grau dos genes e o betweenness considerando os miRNAs e piRNAs. c) Projeção da rede ncRNA com foco nos genes. Arestas mais espessas e vermelhas indicam que o pares de genes compartilham a maior proporção de elementos regulatórios, ou seja o número de vizinhos em comun. Os nós em verde indicam genes de impacto clínico.

Considerando as conexões entre genes e miRNAs, os indices de centralidade apresentaram alta correlação de Pearson (R = 0.83, p-value = 2.2e-16). Por outro lado, considerando apenas as interações entre os genes e piRNAs não foi identificada relações de

linearidade nos dados. O *scatter plot* para as distribuições de grau e betweenness estão representados na Figura 2B. Essas características são importantes para identificar os nós com centralidade alta, uma vez que a disposição dos elementos regulatórios na rede de um processo biológico podem provocar distúrbios nas vias de regulação, impedindo ou mesmo alterando a regulação entre os diversos elementos. No caso da alta correlação na rede de miRNAs, tanto o grau quando o betweeness podem informar os elementros cruciais de uma via biológica. A correlação nula entre o grau e o betweenness na rede de piRNAs pode ser explicada pela biogenese dos piRNAs, uma vez que o número de genes que produz o memso piRNA é baixo, tendo como consequencia uma rede esparsa.

Na rede de miRNAs, o *hsa-miR-155* é o miRNA com o maior número de conexões, com 137 genes-alvo. A análise funcional destes genes apontou enriquecimento para as vias de câncer, especialmente o câncer coloretal. Nós realizamos o enriquecimento funcional para os dez genes com maior número de conexões (*PTEN*, *CDKN1A*, *BCL2*, *CCND1*, *IGF1R*, *CDK6*, *MYC*, *STAT3*, *HHMGA2* e *VEGF*) e encontramos enriquecimento para vias de miRNAs no câncer.

Na rede de piRNAs, o *piR-hsa-2100* apresentou a maior quantidade de conexões, sendo produzido por 581 genes diferentes. Apesar de ser produzido por vários genes diferentes, este piRNA possui suas funções completamente desconhecidas, apontando para a necessidade de mais estudos. Os genes de maior número de conexões foram *FAM225A* e *FAM225B*, que produzem 639 e 630 piRNAs, respectivamente, dos quais 628 são compartilhados entre os dois. Suas conexões são destaque na rede de sobreposição representada na Figura 2C.

Na rede de circRNA, todos eles apresentam apenas uma conexão. O gene *BIRC6* apresentou o maior número de conexões, uma vez que produz 444 isoformas diferentes de circRNAs. O enriquecimento funcional apontou que este gene está relacionado com várias doenças, incluindo câncer de pulmão e hepatite B.

A integração das três redes de ncRNAs permitiu a identificação de 52 genes que apresentam a maior quantidade de ncRNAs relacionados (Figura 2C). Dentre estes genes, oito (*BCL2*, *CCND1*, *GOLGA8B*, *LOC399744*, *LRRC37A2*, *POM121L9P*, *PTEN* e *SPATA31D1* - Figura 2C - genes em verde) foram encontrados nos dados de GWAS em associação à doenças complexas, tais como câncer de próstata, de mama, leucemia, melanoma, diabetes tipo 2 e esquizofrenia.

5. Conclusões

O presente estudo permitiu a identificação de potenciais novos biomarcadores ainda não descritos previamente em associação à doenças complexas. Dentre eles, destacam-se o *piR-hsa-2100* e os genes *FAM225A* e *FAM225B*, que apresentam um grande impacto na rede de piRNA e que possuem suas funções ainda totalmente desconhecidas.

Além disso, a integração das redes permitiu a identificação de um conjunto de genes que compartilham uma grande quantidade de ncRNAs (vizinhos comuns). Esta característica indica o envolvimento simultâneo e dinâmico das três classes de ncRNAs. Estes genes foram encontrados em associação a diversos tipos diferentes de câncer, indicando o seu potencial como alvos terapêuticos ou como biomarcadores de risco.

Os resultados encontrados abrem uma nova perspectiva de estudos para a

investigação e validação dos novos miRNAs, piRNAs, circRNAs e genes encontrados, demonstrando a importante aplicabilidade das redes complexas na análise de dados biológicos para priorização e validação de elementos regulatórios. Tendo em vista os resultados deste trabalho, está em andamento a criação de uma ferramenta para visualização integradas das redes de ncRNAs e e extração de informações biológicos em escala genômica. Além da construção da ferramenta, os trabalhos futuros incluem a aplicação de novas métricas de redes complexas para análises de agrupamento e redundância em redes regulatórias.

Referências

Bahn, J. H., Zhang, Q., Li, F., Chan, T.-M., Lin, X., Kim, Y., Wong, D. T., and Xiao, X. (2015). The landscape of microrna, piwi-interacting rna, and circular rna in human saliva. *Clinical chemistry*, 61(1):221–230.

Barthelemy, M. (2018). *Morphogenesis of spatial networks*. Springer.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2018). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Burgos, E., Ceva, H., Hernández, L., Perazzo, R. P., Devoto, M., and Medan, D. (2008). Two classes of bipartite networks: nested biological and social systems. *Physical Review E*, 78(4):046113.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., Huang, W.-C., Sun, T.-H., Tu, S.-J., Lee, W.-H., et al. (2017). mirtarbase update 2018: a resource for experimentally validated microrna-target interactions. *Nucleic acids research*, 46(D1):D296–D302.

Costa, F. F. (2010). Non-coding rnas: meet thy masters. *Bioessays*, 32(7):599–608.

Esteller, M. (2011). Non-coding rnas in human disease. *Nature Reviews Genetics*, 12(12):861.

Fan, Y., Habib, M., and Xia, J. (2018). Xeno-mirnet: a comprehensive database and analytics platform to explore xeno-mirnas and their potential targets. *PeerJ*, 6:e5650.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circbase: a database for circular rnas. *Rna*, 20(11):1666–1670.

Kontou, P. I., Pavlopoulou, A., Dimou, N. L., Pavlopoulos, G. A., and Bagos, P. G. (2016). Network analysis of genes and their association with diseases. *Gene*, 590(1):68–78.

Maegdefessel, L. (2014). The emerging role of micro rna s in cardiovascular disease. *Journal of internal medicine*, 276(6):633–644.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory rna. *Nature Reviews Genetics*, 15(6):423.

Nair, V. S., Pritchard, C. C., Tewari, M., and Ioannidis, J. P. (2014). Design and analysis for studying micrornas in human disease: a primer on-omic technologies. *American journal of epidemiology*, 180(2):140–152.

Rasool, M., Malik, A., Zahid, S., Ashraf, M. A. B., Qazi, M. H., Asif, M., Zaheer, A., Arshad, M., Raza, A., and Jamal, M. S. (2016). Non-coding rnas in cancer diagnosis and therapy. *Non-coding RNA research*, 1(1):69–76.

Shankaraiah, R. C., Veronese, A., Sabbioni, S., and Negrini, M. (2018). Non-coding rnas in the reprogramming of glucose metabolism in cancer. *Cancer letters*, 419:167–174.

Vidal, A. F., Sandoval, G. T., Magalhães, L., Santos, S. E., and Ribeiro-dos Santos, Â. (2016). Circular rnas as a new field in gene regulation and their implications in translational research. *Epigenomics*, 8(4):551–562.

Volinia, S., Galasso, M., Costinean, S., Tagliavini, L., Gamberoni, G., Drusco, A., Marchesini, J., Mascellani, N., Sana, M. E., Jarour, R. A., et al. (2010). Reprogramming of mirna networks in cancer and leukemia. *Genome research*, 20(5):589–599.

Wang, J., Zhang, P., Lu, Y., Li, Y., Zheng, Y., Kan, Y., Chen, R., and He, S. (2018). pirbase: a comprehensive database of pirna sequences. *Nucleic acids research*, 47(D1):D175–D180.