https://eventos.utfpr.edu.br//sicite/sicite2020

Estudo de medidas de similaridade entre grafos para construção de árvores filogenéticas

Study of similarity measures between graphs for the construction of phylogenetic trees

Erika Hamakami
erikahamakami@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Fabrício Martins Lopes
fabricio@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

RESUMO

A bioinformática é uma área de pesquisas interdisciplinar que apresenta vários importantes desafios computacionais. Um desses desafios consiste na inferência de árvores filogenéticas, especialmente quando considerados organismos muitos próximos em termos de evolução. Para a construção de árvores filogenéticas é preciso analisar diversos aspectos de um ser vivo, dentre eles são analisadas as sequências genéticas que possuem em sua estrutura um caráter conservador, mas também que permita avaliar a sua evolução. Esta pesquisa apresenta uma metodologia para a análise e classificação de sequências de mRNA e lncRNA de diferentes espécies utilizando como base a teoria de redes complexas e reconhecimento de padrões. Mais especificamente, é apresentado como medidas topológicas de redes complexas podem ser usadas como características para a classificação de sequência de transcritos e como tais medidas contribuem para a construção de árvores filogenéticas, indicando resultados promissores.

PALAVRAS-CHAVE: Rede complexa. Bioinformática. Reconhecimento de padrões.

ABSTRACT

Bioinformatics is an interdisciplinary research area that presents several important computational challenges. One of these challenges is the inference of phylogenetic trees, especially when they are considered organisms that are very close in terms of evolution. To construct phylogenetic trees, it is necessary to analyze several aspects of a living being, among them are analyzed the genetic sequences that have in their structure a conservative feature, but also that allow evaluating their evolution. This research presents a methodology for the analysis and classification of mRNA and lncRNA sequences of different species using the complex networks theory and pattern recognition. More specifically, it is presented as topological measurements of complex networks can be used as features for transcript sequence classification and as such measures contribute to the construction of phylogenetic trees, indicating promising results.

KEYWORDS: Complex networks. Bioinformatics. Pattern recognition.

INTRODUÇÃO

Um dos desafios encontrados na Biologia é a construção de árvores filogenéticas que represente o grau de ancestralidade entre os seres vivos da melhor maneira. Este desafio deve-se à grande quantidade de informações e cálculos exigidos para resultar em uma árvore evolutiva com acurácia. Por isso, a computação possui papel importante nesta área, auxiliando no tratamento e análise desse grande volume de dados levando a descoberta de conhecimento.

Para este presente trabalho, são analisadas sequências de mRNA e lncRNA encontradas em diferentes espécies de seres vivos, com intuito de analisar os grafos formados pelas sequências e suas similaridades, utilizando redes complexas e medidas de similaridade entre grafos, em conjunto com a ferramenta BASiNET (ITO, E. A., 2018), já que esta serviu como base para que a pesquisa fosse possível. No desenvolvimento deste artigo serão abordados brevemente sobre alguns conceitos básico para melhor compreensão da pesquisa.

A **biologia molecular** consiste no estudo das células com foco na genética e nas interações bioquímicas, a nível molecular, a partir da relação entre DNA, RNA e da síntese proteica. O ácido desoxirribonucleico (DNA) e o ácido ribonucleico (RNA) são cadeias poliméricas responsáveis pelas instruções, transmissões e pelo armazenamento de informações genéticas. (NELSON, D. L., 2014)

Os DNAs são longas cadeias poliméricas que armazenam informações hereditárias na forma de moléculas. A fita simples de DNA é constituída por vários nucleotídeos, onde cada nucleotídeo é formado por uma molécula de desoxirribose (açúcar) e fosfato, ligado à uma base nitrogenada, sendo elas: a adenina (A), timina (T), citosina (C) ou guanina (G). A fita simples, através do pareamento entre bases, forma uma fita complementar a ela, então, A irá parear com a T e a C com a G. Assim, tem-se duas fitas simples de DNA conectadas, cada uma das fitas pode funcionar como fita molde para a síntese de fitas de DNA, ou seja, replicação de DNA (ALBERTS, B., 2010).

Para que o DNA armazene as informações hereditárias é necessário que ele cumpra com outros processos. Então, ele irá atuar em conjunto com o ácido ribonucleico, RNA, que também é um polímero, porém constituído pelo açúcar ribose, um grupo fosfato e possui como base uracila (U), adenina (A), citosina (C) ou guanina (G). No processo chamado de transcrição, um molde da fita de DNA será usado para que uma fita de RNA seja gerada, então cada base nitrogenada do DNA será ligada à uma base do RNA, então respectivamente, A, T, C, G (DNA) com U, A, G, C (RNA). Em seguida, temos o processo de tradução, onde cada fita de RNA ou RNA mensageiro (mRNA), pois atua como um intermediário, tem seus códons (grupo de três nucleotídeos que corresponde à um aminoácido) lidos. Então o RNA transportador (tRNA) transporta os aminoácidos guiando-os para a síntese proteica conforme as informações contidas no DNA (ALBERTS, B., 2010).

A **árvore filogenética** foi criada para tornar possível o conhecimento das características que convergiram ou divergiram entre os seres vivos de acordo com a evolução (NELSON, D. L., 2014).

Ela é separada por grupos, onde todas possuem um ancestral em comum e com a evolução houveram três ramificações de seres vivos: as bactérias, as arqueobactérias e os eucariotos (NELSON, D. L., 2014). Dentre eles, os procariotos

são os que apresentam maior desafio em relação à classificação de todas as espécies deste grupo. Atualmente, existem 99% de espécies de procariotos que não foram caracterizados ou classificados. Para acelerar e otimizar o processo de classificação das espécies são criados métodos computacionais que, a partir da sequência de DNA, organizam os genes e comparam entre os seres, a fim de obter o número de diferenças existentes permitindo um resultado direto, objetivo e qualitativo sobre a distância evolutiva dos seres envolvidos (ALBERTS, B., 2010).

A evolução dos seres vivos deve-se ao processo de mutação, ou seja, erros na sequência de nucleotídeos. Os segmentos de DNA que não possuem funções fundamentais, sofrem mudanças com mais frequência, porém existem os genes chamados de altamente conservados. Estes genes dificilmente sobrevivem quando sofrem algum tipo de mutação, mantendo-se o mesmo em várias espécies, desta forma, temos um segmento do DNA que permanece com as mesmas características em todas as espécies. Neste caso é analisado o 16S rRNA que está localizado na menor subunidade (30S) de um ribossomo, ele é um componente do rRNA, que resulta em informações válidas sobre a distância evolutiva, já que durante o processo de tradução, crucial em todos os seres vivos, o rRNA é conservado (ALBERTS, B., 2010).

Nesta mesma linha de pesquisa sobre genes e seu nível de conservação ao longo da evolução, têm-se os estudos relacionados aos RNAs não codificante (ncRNA) que são uma classe de moléculas de RNA que não são traduzidas como proteína, dentre elas temos os ncRNAs maiores que 200 nucleotídeos, como o rRNA e o *long noncoding RNA* (lncRNA) e os ncRNAs entre 40 e 200 nucleotídeos, os *small noncoding* RNA (sncRNA), como o tRNA (ENG, Y., 2019). Para esta pesquisa é interessante mencionar os lncRNAs, que apresentam influência sobre o processo de desenvolvimento de doenças, como o câncer, mas ainda não possuem suas funções bem definidas. Porém, têm sido de interesse para os pesquisadores analisá-las e identificar se possuem regiões na sua estrutura e se estas são conservadas ao logo da evolução dos seres vivos (KUNG, J. T, 2013).

Como mencionado, ainda restam 99\% da população bacteriana a serem classificadas. Assim, são criados diversos métodos de classificação de seres vivos e sequências genéticas. Destes métodos, temos Hibridização DNA-DNA (DDH), *Average nucleotide identity* (ANI), *Multilocus sequence analysis* (MLSA) e o BASiNET. Para este trabalho daremos foco ao BASiNET, que trabalha com redes complexas para a análise de sequências genéticas.

Uma rede é um grafo G composto por um grupo de vértices V(G) conectados através de arestas E(G). As arestas determinam algum tipo de vínculo entre um par de vértices, considerando o problema encontrado, podendo receber pesos. Um grafo pode ser direcionado ou não, ou seja, as arestas podem ter sentido único ou duplo. Se G é direcionado, então possui uma aresta que parte de um vértice origem à um vértice destino, não podendo realizar o caminho contrário, mas podem realizar o caminho para ele mesmo (SILVA CARVALHO, H. D. da, 2016). Uma rede pode ser complexa, possibilitando que sejam extraídas diversas medidas para caracterizá-la, dentre elas temos 10 medidas topológicas trabalhadas pelo BASiNET e por esta pesquisa: *assortativity* (ASS), *average degree* (DEG), *maximum degree* (MAX), *minimum degree* (MIN), *average betweenness centrality* (BET), *clustering coefficient* (CC), *average short path length* (ASPL), *average standard deviation* (SD),

frequency of motifs de tamanho 3 (MT3) e *frequency of motifs* com tamanho 4 (MT4) (ITO, E. A., 2018).

MATERIAL E MÉTODOS

Este trabalho foi desenvolvido utilizando o software R e construído com base na ferramenta BASiNET. O R é um software de análise, manipulação e visualização de dados disponível gratuitamente. Neste projeto, foram trabalhados com sequências no formato FASTA de 10 sequências de mRNA e 11 sequências de lncRNA obtidos do pacote BASiNET.

Em conjunto com as algumas linhas de código e funções do BASiNET (***createNet***, ***threshold*** e ***measures***), criou-se uma função ***tree***, que aproveita parte do código encontrado na função ***classification*** do BASiNET e implementou-se novas linhas de código para que o algoritmo retornasse dados a serem analisados e que resultem em medidas de similaridade entre grafos, sendo visualizado através de um dendrograma.

O BASiNET trabalha com redes complexas representadas por grafos G não direcionados, sendo cada vértice V ligado por arestas ou caminhos C, G = {V, C}. Cada caminho é representado por (i, j), ou seja, é uma aresta que liga o vértice i ao vértice j, independente da direção (ITO, E. A., 2018).

Uma rede é gerada a partir da função *createNet*. Nesta função, a variável auxiliar irá receber uma palavra inteira *word* = 3, isto é, 3 nucleotídeos adjacentes, de uma sequência FASTA, assim forma-se um vértice e a cada dois vértices formados é somado um *step* = 1 ao valor de ligação entres eles (ITO, E. A., 2018). Então, o grafo construído é representado por uma matriz adjacente A, obtida através da função *threshold*. Onde um valor X de corte será aplicado aos vértices, cada vértice que apresentar valores menores que X receberão valor 0. Então, W é o peso da matriz e $w_{i,j}$ = w(i,j). Se X é o valor limitante, então, $a_{i,j}$ = 1 se $w_{i,j}$ > X, e $a_{i,j}$= 0, caso contrário. Isso permite obter a ocorrência de ligações entre vértices, assim, após ser feita cada iteração são considerados aqueles com maior frequência (ITO, E. A., 2018). Mas neste trabalho, para a análise de dados, o *threshold* foi limitado de 200 para fazer até 30 níveis.

A função *measures* foi utilizada para devolver medidas topológicas importantes durante a análise da rede e sua caracterização, representação, classificação e modelagem. A função retorna um vetor com 10 medidas da sequência analisada, conforme as medidas apresentadas na Introdução. A diferença entre as medidas da sequência A com a sequência B, pode representar a distância deles dentro de uma árvore filogenética, quanto menor a diferença, maior é sua similaridade. E para cada sequência analisada foram calculadas as medidas topológicas antes e depois do *thrsehold*. Essas foram inseridas em um *dataframe*, conforme a análise das sequências, então, cada linha representa uma sequência diferente e cada coluna apresenta os valores das medidas, conforme a Figura 1. A partir disso, o *dataframe* foi passado como parâmetro dentro de uma função do R, dist(). Esta função irá calcular a distância entre duas sequências, p e q, a partir das medidas topológicas, se p = (p_1, p_2, ..., p_n) e q = (q_1, q_2, ..., q_n), então, distância euclidiana conforme Eq. (1) (ANTON, H., 2014), devolvendo uma matriz

quadrada que possui sequência na linha e na coluna e preenchida com os valores de distâncias entre cada sequência.

$$d(p,q) = d(p,q) =$$
$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_i - p_i)^2} = \sqrt{\sum_{i=1}^{n} = (q_i - p_i)^2} \quad (1)$$

Então, foi realizado um cluster com o resultado das distâncias entre as sequências utilizando a função do R, hclust(), e para conseguirmos analisar visualmente a relação de similaridade entre as sequências, com um *plot* obtém-se um dendrograma, ou seja, um diagrama de árvore, que mostra o agrupamento das sequências mRNA e lncRNA e seus níveis de similaridade de acordo com a matriz fornecida pela função dist().

RESULTADOS E DISCUSSÃO

Esta pesquisa faz a tentativa de analisar como sequências de mRNA e lncRNA de diferentes espécies podem apresentar resultados que contribuam para a pesquisa e construção de árvores filogenéticas.

Após a implementação da função ***tree***, 21 sequências genômicas de espécies diferentes são analisadas. As sequências são analisadas como redes complexas biológicas, então cada uma delas é representada por um grafo. Através do grafo é possível fazer a extração das medidas topológicas citadas na metodologia. Essas características genômicas são organizadas em um *dataframe*, onde as linhas representam cada sequência de mRNA e lncRNA e as colunas todas as medidas extraídas, sendo elas antes e depois do *threshold* ser aplicado, conforme Figura 1.

Então, a partir do *dataframe* criado é possível calcular a distância euclidiana entre as sequências. A cada duas sequências comparadas resulta em um valor que permite relacionar a similaridade que uma sequência possui em relação a outra, de acordo com a Figura 2.

E como podemos observar na Figura 3, é possível comparar as sequências através das medidas e características extraídas dos grafos gerados pelas sequências, permitindo a construção de uma árvore que relaciona a proximidade entre as sequências de mRNA e lncRNA. Nota-se que as sequências foram agrupadas em alguns níveis diferentes, sendo que os lncRNA permaneceram em grupos próximos e sem dispersão para outros níveis. Por outro lado, as sequências de mRNA formam grupos em níveis diferentes, sendo possível observar que o mRNA 5, apresentou ter mais similaridade com lncRNAs do que com aqueles da mesma classe.

Figura 1 – Parte do *dataframe* com medidas topológicas retiradas do grafo antes de aplicar o *threshold* e após o primeiro *threshold* (t1)

	ASPL	CC	DEG	ASS	BET	SD	MAX	MIN	MT3	MT4
mRNA 1	1.609623	0.4353731	36.781250	0.192745301	19.20312	15.534364	5	50	14266	240794
mRNA 2	1.545139	0.5842933	66.218750	0.008496231	17.17188	51.736656	24	63	18866	313362
mRNA 3	1.668155	0.5122894	42.093750	0.081596898	21.04688	29.715967	15	53	12779	198921
mRNA 4	1.432044	0.6612256	70.968750	-0.021215599	13.60937	33.309024	18	1	24391	414073
mRNA 5	2.016393	0.2535627	13.451613	0.191529214	31.00000	9.874088	15	19	3382	38061
mRNA 6	1.414683	0.6432304	73.718750	0.131217219	13.06250	31.820164	10	54	25574	443580
mRNA 7	1.478175	0.6361557	75.562500	0.130121126	15.06250	46.574356	7	47	21656	362042
mRNA 8	1.482639	0.5791078	57.218750	0.039864609	15.20312	24.160488	4	54	21757	377471
mRNA 9	1.371528	0.7111466	97.875000	0.047951320	11.70312	54.013667	16	58	27643	464549
mRNA 10	1.445437	0.6421586	70.562500	0.024273456	14.03125	37.124362	10	57	23685	404788
lncRNA 1	2.139394	0.2595530	8.622222	-0.125071390	25.06667	8.977975	3	7	1147	9350
lncRNA 2	2.179661	0.1986063	9.300000	0.062708616	34.80000	4.854842	7	10	1992	18428
lncRNA 3	2.192012	0.1770186	8.793651	0.040921687	36.95238	4.782895	30	21	1988	18312
lncRNA 4	2.218126	0.1740959	9.301587	0.046011554	37.76190	5.075669	10	9	2102	19949
lncRNA 5	2.073446	0.2061361	10.800000	0.004236167	31.66667	5.703998	15	18	2699	28343
lncRNA 6	2.298515	0.1818645	8.603175	0.065471383	40.25397	4.890873	28	31	1725	14930
lncRNA 7	2.163944	0.1939421	10.137931	0.074558931	33.17241	5.494649	47	24	1926	17484
lncRNA 8	2.144897	0.1750480	10.193548	-0.004525474	34.91935	6.224854	56	11	2301	22861
lncRNA 9	2.230037	0.2050406	9.161290	-0.009732208	37.51613	5.295199	14	26	2021	18733
lncRNA 10	2.297491	0.1756026	8.444444	-0.054018168	40.22222	4.848604	8	5	1795	15783
lncRNA 11	2.133898	0.2030325	10.233333	-0.037005587	33.45000	6.344267	4	35	2338	23365

	ASPL_t1	CC_t1	DEG_t1	ASS_t1	BET_t1	SD_t1	MAX_t1	MIN_t1	MT3_t1	MT4_t1
mRNA 1	1.609623	0.4353731	36.781250	0.192745301	19.20312	15.534364	5	50	14266	240794
mRNA 2	1.545139	0.5842933	66.218750	0.008496231	17.17188	51.736656	24	63	18866	313362
mRNA 3	1.668155	0.5122894	42.093750	0.081596898	21.04688	29.715967	15	53	12779	198921
mRNA 4	1.432044	0.6612256	70.968750	-0.021215599	13.60937	33.309024	18	1	24391	414073
mRNA 5	2.016393	0.2535627	13.451613	0.191529214	31.00000	9.874088	15	19	3382	38061
mRNA 6	1.414683	0.6432304	73.718750	0.131217219	13.06250	31.820164	10	54	25574	443580
mRNA 7	1.478175	0.6361557	75.562500	0.130121126	15.06250	46.574356	7	47	21656	362042
mRNA 8	1.482639	0.5791078	57.218750	0.039864609	15.20312	24.160488	4	54	21757	377471
mRNA 9	1.371528	0.7111466	97.875000	0.047951320	11.70312	54.013667	16	58	27643	464549
mRNA 10	1.445437	0.6421586	70.562500	0.024273456	14.03125	37.124362	10	57	23685	404788
lncRNA 1	2.139394	0.2595530	8.622222	-0.125071390	25.06667	8.977975	3	7	1147	9350
lncRNA 2	2.179661	0.1986063	9.300000	0.062708616	34.80000	4.854842	7	10	1992	18428
lncRNA 3	2.192012	0.1770186	8.793651	0.040921687	36.95238	4.782895	30	21	1988	18312
lncRNA 4	2.218126	0.1740959	9.301587	0.046011554	37.76190	5.075669	10	9	2102	19949
lncRNA 5	2.073446	0.2061361	10.800000	0.004236167	31.66667	5.703998	15	18	2699	28343
lncRNA 6	2.298515	0.1818645	8.603175	0.065471383	40.25397	4.890873	28	31	1725	14930
lncRNA 7	2.163944	0.1939421	10.137931	0.074558931	33.17241	5.494649	47	24	1926	17484
lncRNA 8	2.144897	0.1750480	10.193548	-0.004525474	34.91935	6.224854	56	11	2301	22861
lncRNA 9	2.230037	0.2050406	9.161290	-0.009732208	37.51613	5.295199	14	26	2021	18733
lncRNA 10	2.297491	0.1756026	8.444444	-0.054018168	40.22222	4.848604	8	5	1795	15783
lncRNA 11	2.133898	0.2030325	10.233333	-0.037005587	33.45000	6.344267	4	35	2338	23365

Fonte: Autoria própria (2020).

Figura 2 - Matriz com as distâncias euclidianas calculadas para cada par de sequência a partir das medidas topológicas.

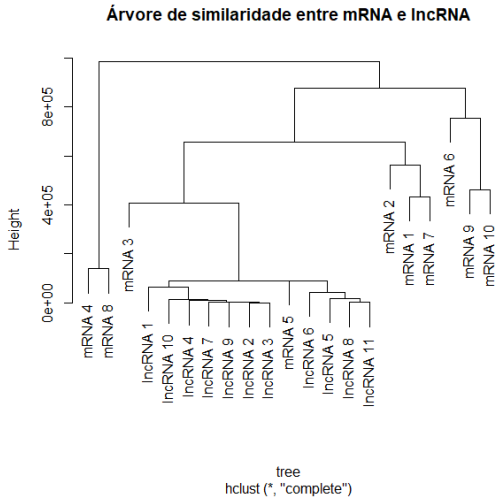
	mRNA 1	mRNA 2	mRNA 3	mRNA 4	mRNA 5	mRNA 6	mRNA 7	mRNA 8	mRNA 9
mRNA 1	0.0000								
mRNA 2	562341.4898	0.0000							
mRNA 3	524282.8806	476771.7277	0.0000						
mRNA 4	468657.8116	854071.4783	884014.4309	0.0000					
mRNA 5	507799.6977	531202.7506	378699.5169	935962.4847	0.0000				
mRNA 6	714677.3854	612001.7784	665813.4321	889854.3769	743169.9530	0.0000			
mRNA 7	434416.6819	517281.1499	544860.5066	635104.3202	619015.6322	595082.4821	0.0000		
mRNA 8	343225.9954	770239.2709	790264.3308	138436.6610	833343.2563	836071.2268	560973.1028	0.0000	
mRNA 9	781894.0042	672698.2532	713736.3895	935226.2951	838856.3744	693984.4528	672358.9759	882086.5874	0.0000
mRNA 10	789138.3615	703180.7104	694450.4999	959666.0406	810125.4033	755591.3808	715762.2893	909141.5156	461681.8329
lncRNA 1	543489.0533	554265.0091	399242.9462	967634.1102	89562.0393	789931.8631	643639.0439	866087.5550	855886.9870
lncRNA 2	543134.4892	551049.4781	404974.4268	981059.4458	65745.0458	793365.0760	654534.4886	875585.0653	874490.3537
lncRNA 3	543254.2586	551216.8111	405041.8612	981206.3262	65794.0186	793542.7661	654632.3459	875725.5055	874624.7700
lncRNA 4	541473.6841	548805.4336	403991.1264	978980.1588	65317.5542	790977.9493	653107.9190	873613.2668	872668.4307
lncRNA 5	500900.2981	550066.6296	398555.3746	941259.5898	89249.2996	795634.5430	628933.9734	831021.8050	851495.9666
lncRNA 6	529589.4367	563760.3943	407576.7514	969670.8854	81973.6436	808789.6282	649335.6485	859956.9267	872179.2141
lncRNA 7	530149.9491	536617.1317	392598.4267	952665.3644	64561.1126	772017.2951	635635.5490	851553.7556	849064.3862
lncRNA 8	508841.0371	550488.3220	397550.9199	941406.1693	82833.5238	792406.5382	630453.3864	836936.0892	850839.6870
lncRNA 9	542772.4243	550580.5672	404742.2170	980604.2632	65624.3621	792868.5745	654209.1814	875155.7647	874106.5965
lncRNA 10	546086.5844	555021.7244	406786.1697	984664.8756	67450.5907	797554.8139	657040.8797	879020.6483	877696.1725
lncRNA 11	500891.7101	538021.1097	387975.0436	927897.3730	81537.3533	775975.9434	617189.6230	824676.5342	834449.5333

	mRNA 10	lncRNA 1	lncRNA 2	lncRNA 3	lncRNA 4	lncRNA 5	lncRNA 6	lncRNA 7	lncRNA 8
mRNA 1									
mRNA 2									
mRNA 3									
mRNA 4									
mRNA 5									
mRNA 6									
mRNA 7									
mRNA 8									
mRNA 9									
mRNA 10	0.0000								
lncRNA 1	819032.9780	0.0000							
lncRNA 2	845832.3230	59519.3517	0.0000						
lncRNA 3	845935.0411	59152.5153	422.7995	0.0000					
lncRNA 4	844356.4401	64317.3707	5232.2566	5629.3906	0.0000				
lncRNA 5	826973.4152	82045.3554	87676.9988	87531.7866	89522.4284	0.0000			
lncRNA 6	845065.2894	43794.6402	63399.7303	63096.1283	67345.2687	41785.3247	0.0000		
lncRNA 7	818844.2748	54648.5511	3314.3425	2914.0138	8490.7033	84469.8682	59494.3258	0.0000	
lncRNA 8	824231.4051	65104.4585	75248.3865	75046.7712	77849.6425	17078.7191	24708.1144	71538.8603	0.0000
lncRNA 9	845502.2088	60462.3000	1074.4547	1459.0808	4179.7227	87975.7043	64140.5916	4324.9171	75702.9005
lncRNA 10	848410.5322	51339.4592	9095.3908	8698.3675	14325.1051	85081.9210	57012.9865	5888.9014	71354.6538
lncRNA 11	804751.7241	65030.9330	74913.6038	74721.3623	77393.4972	15508.5550	26278.4312	70891.8075	1570.6820

	lncRNA 9	lncRNA 10	lncRNA 11
mRNA 1			
mRNA 2			
mRNA 3			
mRNA 4			
mRNA 5			
mRNA 6			
mRNA 7			
mRNA 8			
mRNA 9			
mRNA 10			
lncRNA 1			
lncRNA 2			
lncRNA 3			
lncRNA 4			
lncRNA 5			
lncRNA 6			
lncRNA 7			
lncRNA 8			
lncRNA 9	0.0000		
lncRNA 10	10147.3649	0.0000	
lncRNA 11	75344.2877	71221.0329	0.0000

Fonte: Autoria própria (2020).

Figura 3 – Árvore de similaridade entre as sequências de mRNA e lncRNA



Fonte: Autoria própria (2020).

CONCLUSÕES

O objetivo deste trabalho foi implementar uma função que pudesse analisar diferentes tipos de sequências genômicas, mais especificamente sequências de mRNA e lncRNA de várias espécies, com finalidade de explorar conceitos de redes complexas, teoria dos grafos, bioinformática e reconhecimento de padrões, além de trabalhar com a extração de medidas topológicas das sequências através dos grafos gerados por cada uma delas.

A extração de medidas topológicas por meio da teoria de redes complexas, permitiu caracterizar as sequências e levou a comparação de similaridade entre elas por meio de seus respectivos vetores de características. Dessa forma, os resultados obtidos indicam que é possível utilizar tal metodologia para a inferência de árvore filogenéticas a partir das sequências biológicas. Também foi possível indicar a possibilidade de comparar sequências de mRNA e lncRNA de espécies diferentes, buscando níveis de similaridades, com a finalidade de contribuir para as pesquisas relacionadas à construção e análise de árvores filogenéticas.

Este trabalho apresenta uma primeira iniciativa para a aplicação da teoria de redes complexas e suas medidas na inferência de árvores filogenéticas, apresentando resultados adequados. Como trabalho futuro se espera o desenvolvimento de uma metodologia baseada neste trabalho que seja possível inferir árvores filogenéticas a partir de diferentes organismos.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de Iniciação Científica (Edital PROPPG 2019/2020 - PIBIC) concedida à acadêmica Erika Hamakami.

REFERÊNCIAS

ANTON, H.; RORRES, C. ***Elementary linear algebra: applications version.*** Ed. (11) Estados Unidos: Wiley, 2014. ISBN 9781118434413.

ITO, E. A. et al. ***BASiNET—BiologicAl Sequences NETwork: a case study on coding andnon-coding RNAs identification****. Nucleic acids research*, Oxford University Press, v. 46, n. 16, p.e96–e96, 2018.

ALBERTS, B. **Biologia molecular da célula**. Porto Alegre: Artes Medicas, 2010. ISBN9788536320663.

NELSON, D. L. **Princípios de bioquímica de Lehninger**. Porto Alegre: Artes Medicas,2014. ISBN 9788582710739.

ENG, Y.; LI, J.; ZHU, L. ***Chapter 8 - Cancer and non-coding RNAs***. In:FERGUSON, B. S. (Ed.).Nutritional Epigenomics. [S.l.]: Academic Press, 2019. v. 14.(Translational Epigenetics). P. 119–132.

KUNG, J. T.; COLOGNORI, D.; LEE, J. T. ***Long noncoding RNAs: past, present, andfuture***.Genetics, Genetics Soc America, v. 193, n. 3, p. 651–669, 2013.

SILVA CARVALHO, H. D. da; MARTINS, M. **Teoria dos Grafos e o Problema daGeração de Árvores Filogenéticas**. Repositório Digital-Trabalhos de Conclusão de Curso(Graduação), n. 1, 2016.