UNIVERSIDADE ESTADUAL PAULISTA "JULIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
DEPARTAMENTO DE FÍSICA E BIOFÍSICA
BACHARELADO EM FÍSICA MÉDICA

ANDRÉ LUIZ MOLAN

CRIAÇÃO DE UM ALGORITMO PARA ORGANIZAÇÃO HIERÁRQUICA DE ELEMENTOS INTERAGENTES: UMA APLICAÇÃO EM REDES PROTEICAS, PEQUENAS MOLÉCULAS E miRNAs

Botucatu, SP
Maio 2014

ANDRÉ LUIZ MOLAN

CRIAÇÃO DE UM ALGORITMO PARA ORGANIZAÇÃO HIERÁRQUICA DE ELEMENTOS INTERAGENTES: UMA APLICAÇÃO EM REDES PROTEICAS, PEQUENAS MOLÉCULAS E miRNAs

Trabalho de conclusão de curso apresentado à Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, como parte dos requisitos para a obtenção do título de Bacharel em Física Médica.

Orientador: **Prof. Dr. José Luiz Rybarczyk Filho**

Botucatu, SP
Maio 2014

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE - CRB 8/5651

Molan, André Luiz.
Criação de um algoritmo para organização hierárquica de elementos interagentes : uma aplicação em redes proteicas, pequenas moléculas e miRNAs / André Luiz Molan. - Botucatu, 2014

Trabalho de conclusão de curso (bacharelado - Física Médica) - Universidade Estadual Paulista, Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Capes: 10501045

1. Banco de dados. 2. Analise por conglomerados. 3. Seqüência de nucleotídeos. 4. Algoritmos de computador.

Palavras-chave: Bancos de dados; Cluster; Modularidade; Rede; Reorganização.

Aos meus pais, robustos pilares responsáveis por me manterem em pé, e ao meu orientador, que acreditou em mim mesmo quando o mais sensato seria desistir.

"It looks like a dream (…). Then you suddenly remember it's all real. Then the long march from the rim of the cave to the edge of the cliff where we flung ourselves off and built our wings on the way down quickens to focus." ***Ray Bradbury. The Book Review, Hymn to humanity from the cathedral of high technology. Los Angeles Times, 1979 November 18th.***

RESUMO

A decodificação do Genoma significou um grande passo para o entendimento do funcionamento celular. Inúmeros experimentos de interações proteicas passaram a ser realizados e uma quantidade enorme de dados foi e ainda é gerado. Para tanto, organizações públicas se uniram e criaram bancos de dados como o EMBL (*European Molecular Biology Laboratory*) e o NIH (*National Institute of Health*). Contudo, a forma como os dados estão organizados em tais repositórios pode, eventualmente, não favorecer a extração de informações úteis a um determinado estudo. Posto isso, faz-se necessário o desenvolvimento de ferramentas que sejam capazes de reorganizar as listagens iniciais, facilitando o trabalho do pesquisador no que se refere à identificação e compreensão de processos biológicos.

Existe, atualmente, um método que agrupa genes em uma lista de maneira que a probabilidade de interação entre dois deles é inversamente proporcional à distância em que estes se encontram na lista. Com base neste conceito, desenvolvemos um algoritmo que reorganiza hierarquicamente um dado conjunto de interações, dispondo-o em uma matriz de adjacência simétrica de modo que ocorra a formação de clusters em torno da diagonal principal desta.

Palavras-chave: bancos de dados, reorganização, cluster, modularidade, rede.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

A estrutura do DNA (Ácido Desoxirribonucleico) foi desvendada em 1953 por James Watson e Francis Crick [Watson e Crick 1953]. Em 1962 foram laureados com o Nobel em medicina e fisiologia pelo feito. Tal descoberta possibilitou uma revolução na ciência, visto que é no DNA onde estão todas as informações genéticas de um organismo.

O DEO (Departamento de Energia) e o NIH (Instituto Nacional de Saúde), ambos americanos, iniciaram o chamado Projeto Genoma. Sendo um trabalho em conjunto de vários países, este projeto visava o sequenciamento do código genético de organismos vivos (vegetais, animais, bactérias, entre outros). Através deste estudo vislumbrava-se a criação de mapas físicos de elevada resolução, sequenciamento completo do genoma humano e criação de bancos de dados, feito alcançado e divulgado à imprensa em 14 de abril de 2003.

Em 1998, americanos, japoneses e australianos, além de alguns países europeus, criaram um órgão internacional para gerenciar o sequenciamento do *Homo sapiens*. Esta organização recebeu o nome de HUGO (*Human Genome Organization*) [Crawford 1989]. O propósito do HUGO era o de organizar todo o trabalho de sequenciamento, armazenando, em um banco de dados central (*Genome Database*), todo o conhecimento adquirido. O mapa físico foi finalizado em 2003 e, com isso, o papel do HUGO passou a ser a disseminação de dados funcionais do genoma e fornecimento de diretrizes para as inúmeras aplicações e implicações deste.

No entanto, o sequenciamento do DNA não foi suficiente para a compreensão do funcionamento celular. Para isso, era preciso entender a maneira como as informações contidas no DNA eram traduzidas em proteínas, as quais, de fato, permitiam a ocorrência de uma reação bioquímica no interior da célula.

Com o intuito de melhor compreender este processo, faz-se necessário uma breve explanação. Inicialmente há a produção do RNA (Ácido Ribonucleico), que é uma molécula similar ao DNA. O RNA transcrito é transportado ao citoplasma, onde ocorre a produção de proteínas. Esta etapa segue exatamente as instruções contidas no DNA, ou seja, cada uma das bases (adenina, citosina, guanina e timina) tem sua base complementar no RNA transcrito (adenina, guanina, citosina e uracila) que por sua vez é traduzido em um componente proteico. Cada conjunto de três bases dá lugar a um aminoácido, e cada conjunto de aminoácido a uma proteína. A região no DNA responsável pela codificação de uma proteína recebe o nome de gene [Nelson e Cox 2004].

Embora não tenha sido suficiente para mostrar o real funcionamento celular, a decodificação do Projeto Genoma deu um grande passo nesta direção [Crawford 1989]. É possível encontrar uma gama enorme de dados acerca do funcionamento celular em nível de interações proteína-proteína. A origem destes podem ser resultados de experimentos de alto desempenho ou resultados de alguma pesquisa direcionada a poucas proteínas ou genes [Edgar, Domrachev e Lash 2002]. Informações sobre interações de pequenas moléculas e proteínas podem ser encontradas em profusão. Já a interação entre pequenas moléculas, drogas ou íons, mesmo não sendo medidas por técnicas de microarranjo (*microarray*), é conhecida sua ação sobre um gene ou um grupo de genes que podem ser catalisados, inibidos ou ativados [Arrel e Terzic 2010, Barve, Rodrigues e Wagner 2012, Berger e Iyengar 2009].

Verifica-se, desta forma, o surgimento de uma complexa rede de interações entre genes, proteínas e pequenas moléculas das células, fato que pode ser observado por meio de rotas bioquímicas. Estas rotas são como uma rede de interação entre metabólitos [Kanehisa e Goto 2000, Okuda et al. 2008, Brohée et al. 2008]. Muitos destes metabólitos são codificados a partir do genoma, enquanto outros são compostos sintetizados pela própria célula.

Existe uma grande quantidade de cientistas trabalhando em cima do enorme volume de informações originárias de diversos experimentos. Isso levou a criação de bancos de dados públicos, como o EMBL (*European Molecular Biology Laboratory*) e o NIH, anteriormente já citado. Entretanto, os dados dentro dessas bases estão organizadas de maneira diferente [Szklarczyk et al. 2011, Knox et al. 2011, Kuhn et al. 2012]. Com isso, o principal desafio é o de transformar esta coleção de dados em informações úteis que levem a uma compreensão mais precisa do funcionamento das células e organismos vivos. Através deste entendimento mais acurado, seria possível, então, o desenvolvimento de ferramentas de diagnóstico de doenças cuja origem reside no mau funcionamento do metabolismo celular.

Atualmente existe um método que agrupa genes em uma lista de tal forma que a probabilidade de que dois deles, distantes N posições, estejam relacionados, decai exponencialmente com N. É preciso ressaltar, porém que o agrupamento é realizado independentemente de resultados de experimentos de expressão. Dado isso, o método consegue formar conjuntos de genes associados a um mesmo processo biológico, transformando a lista inicial [Rybarczyk-Filho et al. 2011]. Esta metodologia trabalha apenas com interações proteicas, mas sabe-se, porém, que em toda rota metabólica podem ser encontradas uma ou mais pequenas moléculas/íons, além de, eventualmente, poder receber a ação de uma droga sobre o organismo. Dessa forma, pensando em uma melhor compreensão

do metabolismo, faz-se necessário o desenvolvimento de uma organização hierárquica que utilize diversos tipos de interação de elementos pertencentes à célula.

2 OBJETIVOS

O presente trabalho tem como objetivo a criação de um algoritmo de organização hierárquica de associação proteína-proteína, pequena molécula-proteína, pequena molécula-pequena molécula, regulação transcricional, associação metabólica e microRNA-proteína, com o intuito de, eventualmente, encontrar clusters que possam representar um dado processo biológico.

3 MATERIAIS E MÉTODOS

3.1 REDE INTEGRADA DE INTERAÇÃO DE GENES HUMANOS (INHGI)

A rede INHGI (*Integrated Network of Human Genes Interactions*) [ACENCIO et al. 2013] contém apenas interações que foram experimentalmente verificadas. Ela foi construída partindo do pressuposto de que dois genes, g_1 e g_2, codificando, respectivamente, as proteínas p_1 e p_2, são genes interagentes. Isso só é possível se:

- p_1 e p_2 interagem fisicamente (interação física de proteínas - PPI);
- o fator de transcrição p_1 regula diretamente a transcrição do gene $g_{2;}$
- as enzimas p_1 e p_2 compartilham metabolitos.

Nesta rede, os dados de interações físicas proteína-proteína foram obtidos da versão 1.3 do repositório HIPPIE (*Human Integrated Protein-Protein Interaction rEference*) [SCHAEFER 2012]. Deste, foram selecionadas apenas interações detectadas por técnicas experimentais que receberam um score de 5 ou mais.

Já as interações de regulação transcricional de humanos foram obtidas da atual versão do HTRIdb (*Human Transcriptional Regulation Interaction database*) [BOVOLENTA et al. 2012]. O HTRIdb é um repositório de interações entre fatores transcricionais humanos e seus genes alvos.

3.2 MODELO DE INTERAÇÕES GENES-AMBIENTE (GENVI)

O GENVI (*Gene-Enviroment Interactions model*) foi desenvolvido com o intuito de estudar a natureza multifatorial do autismo de modo a encontrar os processos biológicos predominantes, componentes celulares e funções moleculares relacionadas à desordem afim de elucidar as contribuições genéticas e/ou ambientais da doença [ZEIDÁN-CHULIÁ, et al. 2013].

3.3 BANCO DE DADOS DE INTERAÇÃO PROTEÍNA – PEQUENA MOLÉCULA E ENTRE PEQUENAS MOLÉCULAS

Visando à confecção do banco de interações químico-químico e químico-proteína, o repositório utilizado foi o STITCH [Kuhn et al. 2012] (*http://stitch.embl.de/*). Este pode ser considerado "irmão" do repositório STRING [Szkçarcyk et al. 2011] (*http://string.embl.de/*), uma base de dados que disponibiliza a existência ou não de interações entre produtos gênicos para 1134 organismos. O STITCH contém informações sobre a existência ou não de interações para cerca de 630 organismos.

3.4 BANCO DE DADOS DE miRNA

Para informações sobre miRNA (micro RNA), a base utilizada foi o *StarBase*, versão 2.0 (*http://starbase.sysu.edu.cn/*). Ela foi concebido para a decodificação de redes de interação de lncRNA (*long noncoding RNA*), miRNA, ceRNA (*competing endogenous RNA*), RBP (*RNA-binding proteins*), mRNA de dados de larga escala e amostras de tumores (14 tipos de câncer e mais de 6000 amostras). O *StarBase* oferece ferramentas web de *miRFunction* e *ceRNAFunction* para predizer a função de ncRNA (miRNAs, RNAs longos não codificantes e pseudogenes) e genes de codificação de proteínas de redes regulatórias de miRNA-mediado.

3.5 ALGORITMO DE ORGANIZAÇÃO DE ELEMENTOS INTERAGENTES

O algoritmo foi implementado utilizando a linguagem de programação C++, com compilador Dev C++ em ambiente *Windows* (sistema operacional *Windows 7*). O equipamento utilizado foi um ACER ASPIRE 5750-6802, com processador Intel Core i5 de 2.4GHz e *Turbo Boost* que eleva o *clock* a 3.0GHz, HD de 500GB e memória RAM DDR3 de

3GB. Para a montagem dos gráficos foi usado o OriginPro 8. Já para a visualização espacial das redes, o software escolhido foi o Cytoscape 3.1.0. A escolha da linguagem se deve ao fato de ela ser robusta e apresentar uma velocidade de processamento muito elevada. Inicialmente, o algoritmo parte de uma rede construída a partir dos bancos de dados anteriormente citados. Esta rede consiste em um arquivo texto, com duas colunas, em que o elemento da direita está interagindo com o da esquerda (neste trabalho, o número de elementos interagentes distintos foi de 860 e o número de interações entre eles de 13504). A primeira tarefa realizada foi a extração do conteúdo da rede e armazenamento deste em uma matriz de 13504 linhas por 2 colunas (Matriz Numérica). Nesse processo os elementos são convertidos em números e seus nomes são armazenados em uma matriz de 860 linhas por 2 colunas (Matriz Legenda), de forma que o número contido na matriz numérica possa, posteriormente, ser relacionado ao nome do elemento. Ambas as matrizes são exemplificadas na Figura 1.

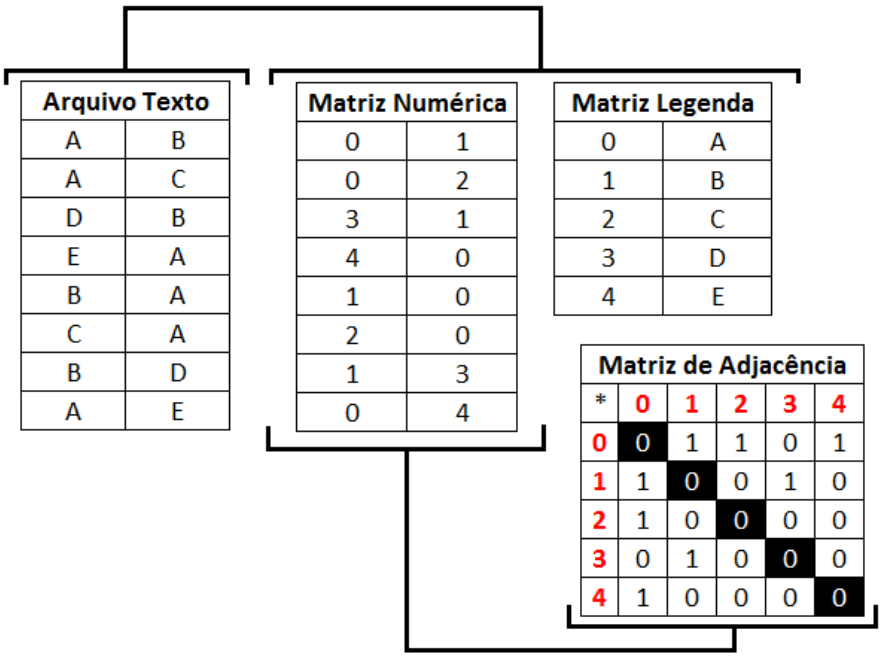

Figura 1 – Exemplificação dos formatos do Arquivo Texto, Matriz Numérica, Matriz Legenda e Matriz de Adjacência e relacionamento entre matrizes e arquivo.

No exemplo da Figura 1, verifica-se que o arquivo texto dá origem às matrizes Numérica e Legenda. A Matriz de Adjacência representa a forma como as ligações entre os elementos estão dispostas. O número 1 indica que há interação, enquanto zero mostra o contrário. Esta matriz, porém, não existe fisicamente. Estamos trabalhando com 860

elementos distintos, o que daria uma matriz 860x860, ou seja, 739600 termos – um volume significativamente elevado. Para evitar isso, trabalhamos somente com as interações, o que reduz a quantidade de termos para 13504 – número inicial de interações. A Matriz Numérica, mais do que armazenar os elementos interagentes em forma de número, armazena os índices da configuração inicial da Matriz de Adjacência. A coluna da esquerda representa a linha, enquanto a da direita, a coluna. Este artifício nada mais é do que a aplicação do conceito de matriz esparsa, que significa ignorar os termos nulos e considerar apenas aqueles preenchidos.

O próximo passo consiste no "embaralhamento" configuração inicial. São selecionados aleatoriamente dois números, variando entre 0 e o número de elementos distintos (no caso, 860). Dado isso, trocamos de lugar as respectivas linhas e colunas. A quantidade de trocas realizadas foi de 10^6. Este processo se faz necessário pois aumenta a eficiência do algoritmo durante o cálculo de energia (ao utilizar o termo energia, entenda grau de proximidade das interações à diagonal principal). Vale ressaltar que em momento algum as interações são destruídas. A matriz é simétrica. Apenas as posições das ligações dentro da Matriz de Adjacência são trocadas. Para cada troca, as matrizes Numérica e Legenda são sempre atualizadas, visto que as posições iniciais mudaram. Com o "embaralhamento" concluído, dá-se início ao cálculo de energia (E), o qual é feito por meio da seguinte fórmula:

$$E = \sum_{i=1}^{N}\sum_{j=1}^{N} d_{i,j}[|a_{i,j}-a_{i+1,j}|+|a_{i,j}-a_{i+1,j+1}|+|a_{i,j}-a_{i-1,j+1}|+|a_{i,j}-a_{i-1,j-1}|+|a_{i,j}-a_{i+1,j-1}|+|a_{i,j}-a_{i-1,j}|+|a_{i,j}-a_{i,j+1}|+|a_{i,j}-a_{i,j-1}|] \quad (1)$$

Nesta fórmula, a_{ij} se refere à posição do elemento alvo na Matriz de Adjacência, enquanto d_{ij} indica a distância (módulo da diferença entre i e j) do elemento alvo para a diagonal principal da Matriz de Adjacência. Por meio desta fórmula, fazemos a checagem dos 8 termos adjacentes a uma dada interação (elemento alvo), conforme mostra a Figura 2.

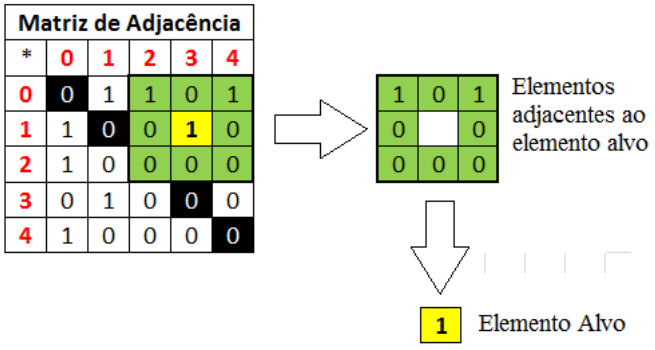

Figura 2 – Cálculo da energia do elemento alvo.

A energia do elemento alvo será a soma dos módulos das diferenças entre o elemento alvo e cada um dos oito termos adjacentes multiplicada pela distância do elemento alvo à diagonal principal da Matriz de Adjacência. Este mesmo procedimento é feito para cada um dos termos não nulos acima da diagonal principal da matriz, uma vez que ela é simétrica e o cálculo abaixo da diagonal seria idêntico. Ao término do cálculo individual, efetua-se o somatório da energia de cada um dos elementos e, então, tem-se a energia total do sistema.

Com o cálculo da energia total inicial realizado, ao algoritmo, agora, seleciona dois números aleatórios, variando entre 0 e o número total de elementos distintos (860), e efetua a troca das respectivas linhas e colunas. Mais uma vez é preciso ressaltar que as interações não são destruídas. Elas são apenas reposicionadas. Feita esta troca, um novo cálculo de energia é realizado. Caso a energia final seja menor que a energia inicial, mantem-se a nova configuração, caso contrário, a configuração inicial é mantida. Independente do que ocorra, uma nova troca de linhas e colunas com base em números aleatórios é realizada e uma nova comparação entre energias é feita. Tal sequência de comparações tem como base o Método de Monte Carlo, o qual tem como principal característica o uso de amostras aleatórias visando à obtenção de resultados numéricos após a realização de uma elevada quantidade de vezes. Aqui, realizamos $2{,}15 \times 10^6$ passos até que se chegasse à configuração final. O objetivo é aproximar os elementos interagentes da diagonal principal da Matriz de Adjacência, aspecto que resultaria na formação de clusters de elementos com uma alta probabilidade de possuírem uma função biológica específica. Ao final da execução são gerados quatro arquivos textos, os quais são usados, posteriormente, para o cálculo de modularidade e plotagem de gráficos para a apresentação e discussão de resultados:

- Matriz Numérica: o primeiro arquivo contém sua configuração inicial (antes do embaralhamento), o segundo sua configuração após o embaralhamento e o terceiro a configuração final após o cálculo de energia.
- Matriz Legenda: um arquivo contendo sua configuração final.

3.6 MODULARIDADE

O cálculo de modularidade é definido como sendo uma janela (a qual consiste em um determinado intervalo) aplicada na lista com os termos interagentes. Ela possui um tamanho mínimo de três e um máximo que pode ser igual ao número de elementos distintos (caso este seja ímpar) ou igual ao número de elementos distintos menos um (caso este seja par). A janela

sempre terá uma quantidade ímpar de termos, uma vez que o módulo dela sempre será atribuído ao elemento central. Para este tipo de cálculo, desenvolvemos um segundo algoritmo. Ele parte do arquivo texto gerado com a configuração final da Matriz Numérica. O tamanho das janelas calculadas se iniciam em 3, indo até 859 (uma vez que estamos trabalhando com 860 elementos distintos). A sequência segue uma progressão aritmética de razão 2, a_1 igual a 3 e a_n igual a 859 (3, 5, 7, 9, ..., 859). O cálculo pode ser exemplificado pela Figura 3.

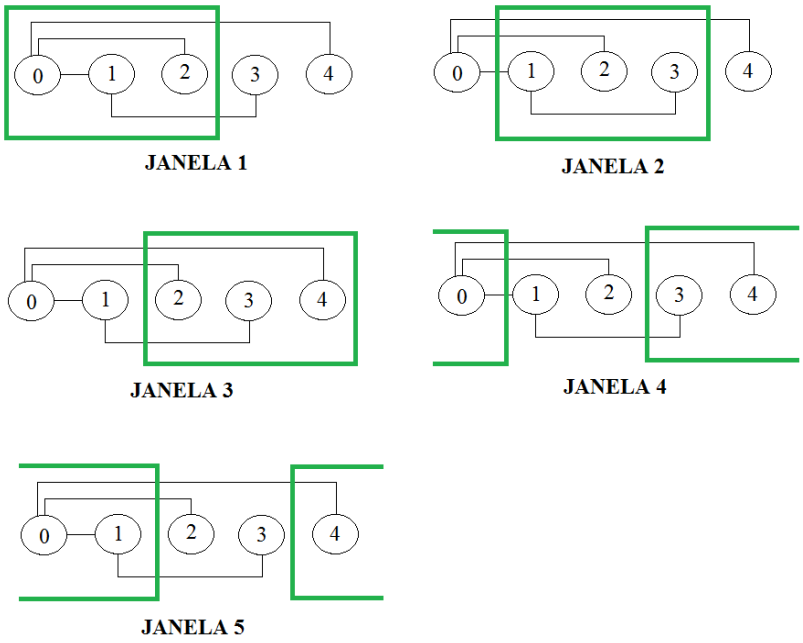

Figura 3 – Exemplo de janelamento para uma janela de tamanho 3 seguindo as interações apresentadas na Figura 1.

O módulo é sempre atribuído ao elemento central de janela, e seu valor é dado pela razão entre o número de ligações que começam e terminam dentro da janela e o número de ligações total da janela. Na Figura 3, seguindo este raciocínio de cálculo, partimos da Janela 1 (para evitar confusões, **tamanho** de janela se refere a quantidade de elementos que ela abrange, enquanto o **número** da janela indica a evolução do janelamento dentro da listagem), cujo elemento central é o termo número 1 e o módulo atribuído a ele é o valor 2/4. Em seguida, na Janela 2, verifica-se que o termo central é o número 2, e o módulo deste é a razão 1/2. O processo se repete para os demais elementos, até que todos tenham sido relacionados a um módulo. No exemplo, o primeiro elemento central encontrado foi o de número 1. Entretanto, todos os termos (0, 1, 2, 3 e 4), em algum momento, são selecionados como

termos centrais e tem seus respectivos módulos calculados, independentemente do tamanho da janela. Pela Janela 5, observa-se que o termo central é o zero, mostrando que ao chegar no elemento 4, voltamos ao início para calcular o módulo dos elementos restantes.

Todas as janelas possíveis foram calculadas (janelas de tamanho 3, 5, 7, ..., 859). O programa gerou um arquivo texto para cada uma delas, perfazendo um total de 429 arquivos (o que corresponde a 429 janelas distintas). A partir daí, 10 janelas foram selecionadas e seu conteúdo plotado em gráficos. Na sequência, das 10 selecionamos 4, e destas, 1, para a montagem da rede final. Este processo se faz necessário para que possamos identificar os clusters, que são representados por picos nestes gráficos. Aqui, a janela escolhida foi a de tamanho 151, apresentando 3 grandes picos principais distintos.

A modularidade é uma característica intrínseca da lista de elementos em função da maneira como esta foi organizada. Diferentes ordenamentos produzem diferentes perfis de ordenamento para janelas de mesmo tamanho. A largura da janela é escolhida de acordo com o interesse do pesquisador. Neste caso, optamos por uma janela que apresentasse pelo menos 3 picos distintos e com módulos elevados.

4 RESULTADOS E DISCUSSÕES

Após a execução do algoritmo de organização hierárquica, utilizando o software OriginPro 8, foram plotados gráficos com base nos dados numéricos (Matriz Numérica). Através da Figura 4, é possível visualizar a disposição dos elementos interagentes em sua configuração inicial, após o embaralhamento e após o cálculo final de energia.

Ao compararmos a configuração final com a pós-embaralhamento, é possível observar, claramente, uma redução sensível do número de interfaces branco-preto e acúmulo das interações ao longo da diagonal principal. É possível, também, verificar a existência de uma série de pequenos agrupamentos, os quais são denominados clusters. Portanto, o algoritmo atingiu o seu propósito, que era o de reorganizar e clusterizar a configuração inicial.

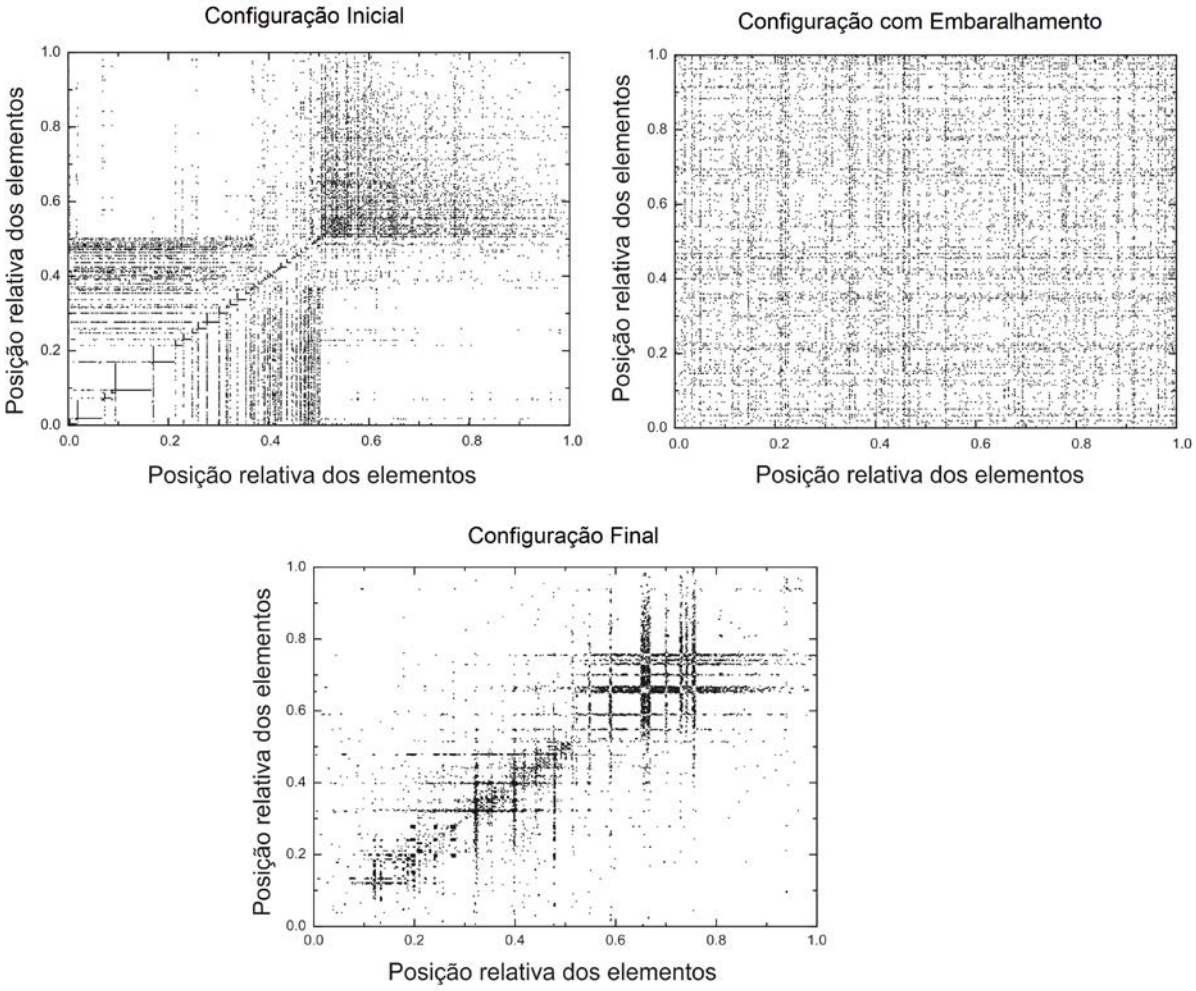

Figura 4 – Disposição gráfica inicial, pós-embaralhamento e pós-cálculo final de energia dos elementos interagentes.

O próximo passo foi a visualização um pouco mais acurada dos clusters formados. É aí que entrou em ação as janelas de modularidades. Foram selecionadas quatro janelas e, por meio do OriginPro 8, foram plotados seus respectivos gráficos, como mostra a Figura 5.

Pela Figura 5, conseguimos observar diferentes configurações para um mesmo conjunto de elementos interagentes em função do tamanho da janela. Nota-se que com o aumento da dimensão do janelamento, ocorre uma redução no número de picos e um aumento no valor dos módulos. Isso ocorre pois, como explicado anteriormente no tópico sobre modularidade, quanto maior a janela, maior a quantidade de ligações, e, portanto, maior o valor do módulo atribuído ao elemento central. Dessa forma, a tendência é uma certa similaridade nos valores obtidos, o que vai suavizando a curva com consequente redução nos picos.

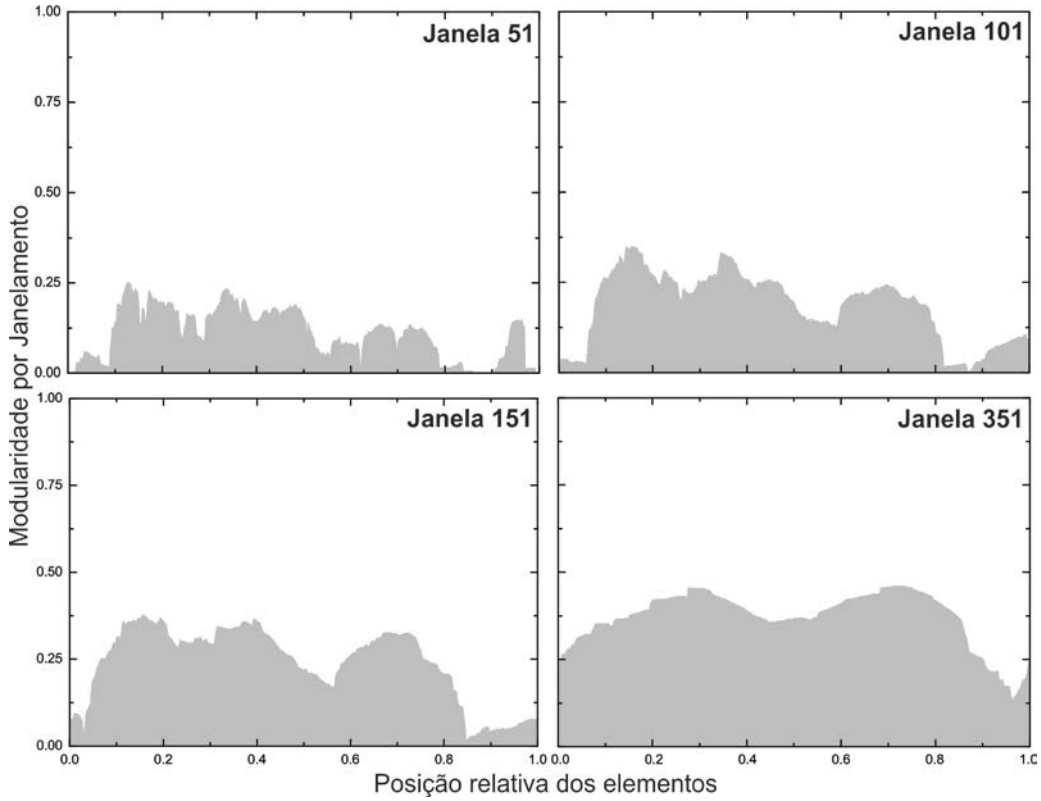

Figura 5 – Janelas de modularidade.

Os picos formados correspondem à formação de clusters. Dessa forma, para melhor observa-los, escolhemos a janela de tamanho 151 e selecionamos três intervalos, com base nos três picos existentes, como ilustra a Figura 6. Esta dimensão de janela foi escolhida somente para efeito didático e exemplificação da formação de clusters. De acordo com o pesquisador, diferentes janelas podem ser selecionadas. Além disso, a janela de tamanho 151 apresenta um módulo elevado e pelo menos 3 grandes picos distintos, como já citado.

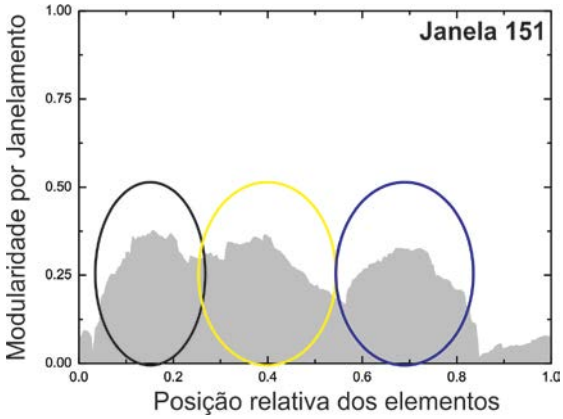

Figura 6 – Seleção de intervalos da janela 151 de acordo com os picos encontrados.

O primeiro intervalo abrange os elementos de 0 a 198. O segundo corresponde aos elementos de 266 a 487. O terceiro, por fim, representa os elementos de 488 a 728. Com o auxilio do software Cytoscape 3.1.0, foram plotadas as redes para cada um dos intervalos, bem como o da rede completa (elementos de 0 a 859), como mostram a Figura 7, Figura 8, Figura 9 e Figura 10, respectivamente.

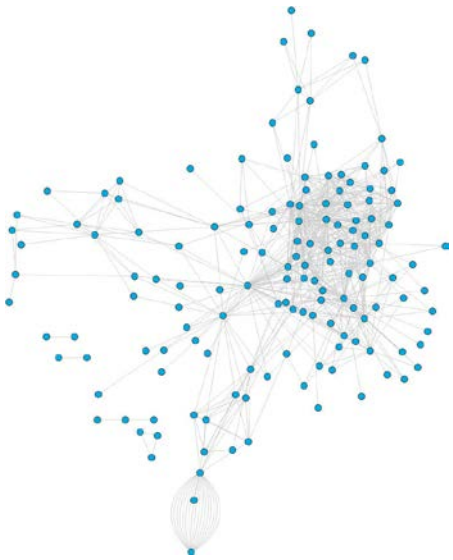

Figura 7 – Rede que compreende os elementos de 0 a 198.

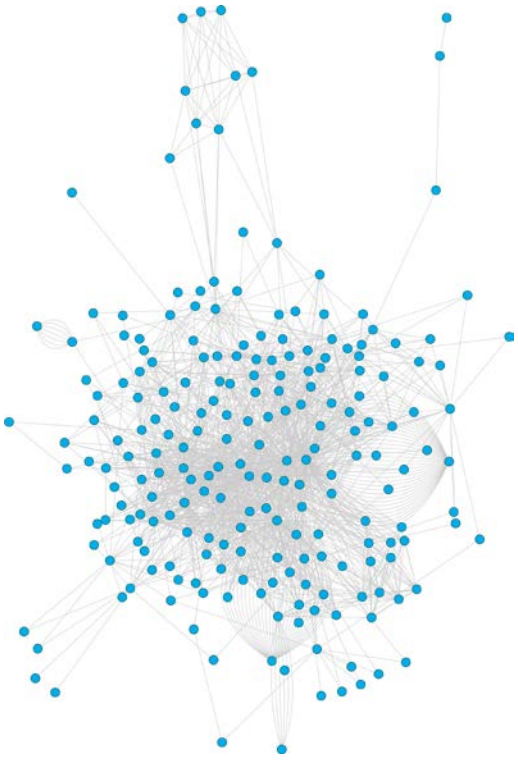

Figura 8 – Rede que compreende os elementos de 266 a 487.

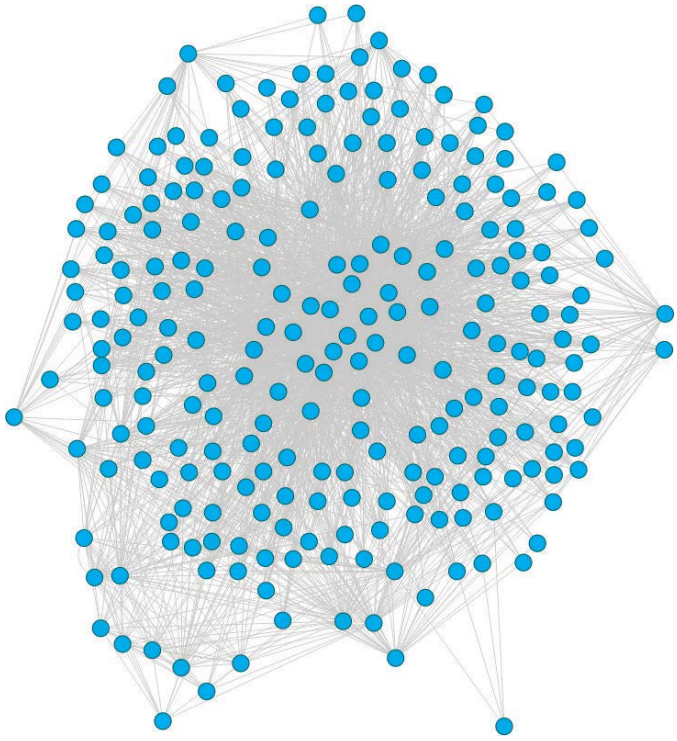

Figura 9 – Rede que compreende os elementos de 488 a 728.

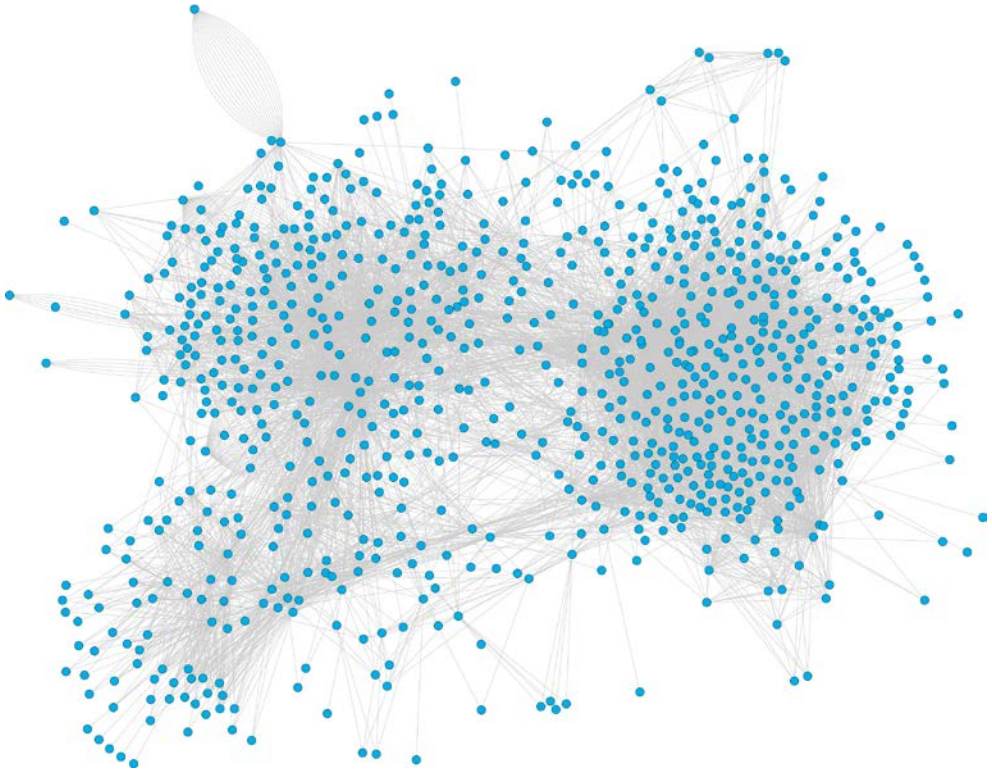

Figura 10 – Rede completa, abrangendo os elementos de 0 a 859.

Para o primeiro intervalo, a quantidade de interações existentes é de 629. No segundo, este número sobe para 2571, enquanto no terceiro cai para 1573, mesmo este intervalo possuindo uma quantidade maior de elementos. Por fim, a rede completa. Esta corresponde a todos os elementos interagentes da rede e apresenta um total de 6745 interações. Inicialmente, quando do cálculo de energia, consideramos todas as ligações, ou seja, A com B e B com A, aspecto que nos proporcionou trabalhar com uma matriz simétrica. Aqui, na montagem gráfica da rede, esta duplicidade é eliminada, isto é, considera-se apenas A com B.

5 CONSIDERAÇÕES FINAIS

Tendo em vista os aspectos analisados, verifica-se que o algoritmo desenvolvido atingiu o seu propósito, que era o de organizar hierarquicamente uma rede de associação proteína-proteína, químico-proteína, químico-químico, regulação transcricional, associação metabólica e microRNA-proteína, permitindo a formação de clusters de elementos que, eventualmente, tivessem algum significado biológico. Tais clusters puderem ser melhor estudados por meio do cálculo das janelas de modularidade e posterior plotagem de determinados intervalos de uma janela de tamanho 151.

O algoritmo se mostrou eficiente, gerando os resultados que esperávamos em um curto espaço de tempo. Da extração inicial dos dados da rede de interações à configuração final após o cálculo de energia, foram necessários 215 minutos. A energia inicial foi reduzida de $8{,}22 \times 10^6$ para $2{,}15 \times 10^6$.

6 PERSPECTIVAS

Para o futuro, melhorias no algoritmo podem ser realizadas. Diferentes vizinhos do elemento alvo podem ser avaliados e uma nova lógica de programação pode ser utilizada. Para reduzir o tempo de processamento poderíamos trabalhar com Álgebra Booleana. O ambiente operacional utilizado poderia ser baseado em UNIX ao invés de Windows. Além disso, o algoritmo poderia ser implementado na versão visual da linguagem C++. Uma outra implementação, ainda, poderia ser o enriquecimento da rede, aspecto este que ampliaria as chances de serem identificados processos biológicos, a integração de informação de expressão gênica, RNA-seq ou SNPs com a rede previamente ordenada. A integração destes dados

podem revelar módulos com altos/baixos níveis de expressão que podem representar uma função essencial para o organismo analisado. Por fim, a criação de serviço online seria interessante. Poderíamos desenvolver um website, onde pesquisadores de outras entidades entrariam apenas com a rede inicial e o software de organização hierárquica a devolveria devidamente reorganizada.

REFERÊNCIAS

ACENCIO ML, BOVOLENTA LA, CAMILO E, LEMKE N (2013) Prediction of Oncogenic Interactions and Cancer-Related Signaling Networks Based on Network Topology. **PLoS ONE 8(10)**: e77521. doi:10.1371/journal.pone.0077521

ARREL, D. K.; TERZIC, A. Network systems biology for drug discovery. **Cli Pharmacol Ther**, v. 88, n.1, p. 120-125, Jul 2010. Disponível em <http://dx.doi.org/10.1038/clpt.2010.91>.

BARVE, A.; RODRIGUES, J. F. M.; WAGNER, A. Superenssential reactions in metabolic networks. **Proc Natl Acad Sci U S A**, v. 109, n. 18, p. E1121-E1130, May 2012. Disponível em: <http://dx.doi.org/10.1073/pnas.1113065109>.

BERGER, S. I.; IYENGAR, R. Network analysis in systems pharmacology. **Bioinformatics**, v. 25, n. 19, p. 2466-2472, Oct 2009. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btp465>.

BOVOLENTA LA, ACENCIO ML, LEMKE N (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. **BMC Genomics 13**: 405. doi: 10.1186/1471-2164-13-405

CRAWFORD, M.H. Hugo: Genome data open do scientists. **Science**, Washinton, USA, v. 216, n. 4937, p. 1565, Dec, 1989.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. **Nucleic Acids Res**, v.30, n. 1, p. 207-210, Jan 2002.

KANEHISA, M.; GOTO, S. Keeg: kyoto encyclopedia of genes and genomes. **Nucleic Acids Res**, v. 28, n. 1, p. 27-30, Jan 2000.

KNOX, c. et al. Drugbank 3.0: a comprehensive resource for ‘omics’ research on drugs. **Nucleic Acids Res**, v. 39, n. Database issue, p. D1035-D1041, Jan 2011. Disponível em: <http://dx.doi.org/10.1093/nar/gkq1126>.

KUHN, M. et al. Stitch 3: zooming in on protein-chemical interactions. **Nucleic Acids Res**, v. 40, n. Database issue, p. D876-D880, Jan 2012. Disponível em: <http://dx.doi.org/10.1093/nar/gkr1011>.

NELSON, D. L.; COX, M. M. **Lehninger Principles of Biochemistry, Fourth Edition**. Fourth Edition. Ed. New York, USA, 2004.

OKUDA, S. et al. Keeg atlas mapping for global analysis of metabolic pathways. **Nucleic Acids Res**, v. 36, n. Web Server issue, p. W423-w426, Jul 2008. Disponível em: <http://dx.doi.org/10.1093/nar/gnk282>.

BROHÉE, S. et al. Network analysis tools: from biological networks to clusters and pahways. **Nat Protoc**, v. 3, n. 10, p. 1616-1629, 2008. Disponível em: <http://dx.doi.org/10.1038/nprot.2008.100>.

RYBARCZYC-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the saccharomyces cerevisiae case. **Nucleic Acids Res**, v. 39, n. 8, p. 3005-3016, Apr 2011. Disponível em <http://dx.doi.org/10.1093/nar/gkq1269>.

SCHAEFER MH, FONTAINE JF, VINAYAGAM A, PORRAS P, WANKER EE, et al. (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. **PLoS One 7**: e31826. doi: 10.1371/journal.pone.0031826

SZKLARCZYK, D. et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Res**, v. 39, n. Database issue, p. D561-568. Disponível em: <http://dx.doi.org/10.1093/nar/gkq973>.

WATSON, J.D.; CRICK F.H. Genetical implications of the structure of deoxyribonucleic acid. **Nature**, London, UK, v.171, n.4361, p.964-967, May, 1953.

WATSON, J.D.; CRICK F.H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. **Nature**, London, UK, v.171, n.4356, p.737-738, Apr, 1953.

ZEIDÁN-CHULIÁ F, RYBARCZYK-FILHO JL, SALMINA AB et al. (2013) Exploring the Multifactorial Nature of Autism Through Computational Systems Biology: Calcium and the Rho GTPase RAC1 Under the Spotlight. **Neuromol Med** (2013) 15:364–383. DOI 10.1007/s12017-013-8224-3