

RESSALVA

Atendendo solicitação do(a)
autor(a), o texto completo desta tese
será disponibilizado somente a partir
de 14/10/2024.

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
INSTITUTO DE BIOCÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

**Desenvolvimento de ferramentas de biologia de sistemas para
avaliação de dados transcriptômicos**

Botucatu, setembro de 2022

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

**Desenvolvimento de ferramentas de biologia de sistemas para
avaliação de dados de transcriptômicos**

Tese de Doutorado apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Doutor no Programa de Pós-Graduação em Biotecnologia

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho

Botucatu, setembro de 2022

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

Pilan, José Rafael.

Desenvolvimento de ferramentas de biologia de sistemas para avaliação de dados de transcriptômicos / José Rafael Pilan. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: José Luiz Rybarczyk Filho

Capes: 90400003

1. Biotecnologia. 2. Computação. 3. Proteínas. 4. Dados de expressão. 5. Biologia.

Palavras-chave: Biotecnologia; Computação; Dados de expressão; Proteínas; Redes.

Agradecimentos

- Ao CNPq por utilizarmos os recursos computacionais referentes aos processos 458810/2013-4 e 473789/2013-2;
- Ao Professor Dr. José Luiz Rybarczyk Filho pela orientação e auxílio;
- À toda minha família, principalmente meu pai José Carlos, minha mãe Maria Aparecida, meu irmão César e minha irmã Elaine pelo apoio e incentivo que foram fundamentais para a conclusão desse trabalho;
- Aos meus filhos Miguel e Marina por serem minha motivação diária em sempre buscar fazer o meu melhor;
- Aos meus amigos André Luiz Molan e Giordano Bruno Sanches Seco por toda a ajuda e motivação;
- À pós-graduação em Biotecnologia do Instituto de Biociências da Unesp de Botucatu e ao seu coordenador Prof. Dr. Valber de Albuquerque Pedrosa representando todos os docentes do programa;
- Aos demais colegas do departamento de Física e Biofísica e do Instituto de Biociências, professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho.

Resumo

A cada dia são desenvolvidos novas metodologias e equipamentos que proporcionam maior facilidade para a elucidação dos RNAs transcritos em situações específicas nos mais diferentes organismos. O que essas metodologias e equipamentos possuem em comum é o enorme volume de dados que podem gerar em cada execução. Devido a grande quantidade de tecnologias que estão surgindo para geração de dados de sequenciamento e expressão gênica precisamos de ferramentas que permitam o estudo e análise integrada dos diferentes dados advindos dos resultados desses experimentos. Nesse trabalho propomos a integração das funções do transcriptograma com uma ferramenta desenvolvida para visualização de dados de expressão sobreposto a uma rede biológica gerando como resultado final um pacote para execução no *layout* de *dashboard* das ferramentas citadas. O pacote foi desenvolvido utilizando a linguagem de programação R em conjunto com as linguagens C e C++ e disponibilizado para o ambiente R. Para validar a ferramenta utilizaremos dados de expressão gênica de amostras de pulmão obtidas das tecnologias de sequenciamento microarranjo e Single-Cell RNA. Como resultados a ferramenta *levi* (*Landscape Expression Visualization Interface*) está disponível para *download* em um dos principais repositórios *online*s de ferramentas de bioinformática, o Bioconductor. A integração das ferramentas do transcriptograma foram adicionadas ao *dashboard* e para validarmos a aplicação realizamos o *workflow* do transcriptograma em dados de microarranjo e RNA-seq de amostras de pessoas com e sem adenocarcinoma de pulmão. Os resultados indicam uma maior acurácia na utilização da ferramenta rTranscriptograma tanto para dados de microarranjo como para dados de RNA-seq.

Abstract

Every day new equipment are developed that provide greater facility for elucidating the RNAs transcribed in specific situations in the most different organisms. What these methodologies and equipment have in common is the large volume of data that they can generate in each execution. Due to the great amount of technologies that are emerging to generate sequencing and gene expression data, we need tools that allow the study and integrated analysis of the different data resulting from the results of these experiments. In this work we propose the integration of the functions of the transcriptogram with a tool to be developed for visualization of expression data superimposed on a biological network, resulting in a package for the execution of the mentioned tools in the final result. The package was developed using the R programming language together with C and C++ languages and made available to the R environment. To validate the tool we will use gene expression data from lung samples obtained from microarray and Single-Cell RNA sequencing Technologies. As a result the levi tool (Landscape Expression Visualization Interface) is available for download in one of the main online repositories of bioinformatics tools, the Bioconductor. The integration of transcriptogram tools has been added to the dashboard and to validate the application, we performed the transcriptogram workflow on microarray data and RNA-seq from samples from people with and without lung adenocarcinoma. The results indicate a greater accuracy in the use of the rTranscriptogram tool both from microarray and RNA-seq data.

Lista de Figuras

2.1	6 "V"s do <i>Big Data</i>	p. 4
2.2	Biologia de Sistemas	p. 5
2.3	Métodos de análises em Biologia de Sistemas	p. 7
2.4	Affymetrix GeneChip - Microarranjo de oligonucleotídeos.	p. 9
2.5	Roche NimbleGen - Microarranjo de oligonucleotídeos.	p. 10
2.6	Agilent - Microarranjo de oligonucleotídeos.	p. 11
2.7	Visão geral de um <i>pipeline</i> para RNA-seq	p. 12
2.8	Tecnologia de sequenciamento de Sanger.	p. 14
2.9	Estratégias de imobilização de <i>templates</i> utilizada pela Applied Biosystems.	p. 15
2.10	Estratégia de amplificação por fase sólida usada pelos equipamentos da Illumina.	p. 16
2.11	Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences.	p. 17
2.12	Isolamento de <i>Single Cell</i> e preparação da biblioteca.	p. 19
2.13	Exemplo de rede direcionada e não-direcionada.	p. 23
2.14	Exemplos de tipos de grafos	p. 23
2.15	Principais modelos de redes	p. 26
4.1	<i>Workflow</i> com as etapas do projeto	p. 31
4.2	<i>Workflow</i> levi	p. 36
4.3	Transformação da rede em matriz de adjacência	p. 38
4.4	Método	p. 39
4.5	Proximidade dos vizinhos em relação ao elemento central na matriz de adjacência	p. 40

4.6	Representação de cinco possíveis distâncias do elemento central em relação a diagonal principal de uma matriz de adjacência.	p. 41
4.7	Avaliação da vizinhança do elemento central no modelo “cruz”.	p. 42
4.8	Gráfico de nível de custo energético por configurações possíveis de uma matriz de adjacência.	p. 43
4.9	Avaliação da vizinhança do elemento central no modelo “X”.	p. 44
4.10	Avaliação da vizinhança do elemento central no modelo “Anel”.	p. 45
4.11	Modularidade por Janela	p. 46
4.12	Resultado da Modularidade por Janela	p. 47
5.1	Rede do String para via de sinalização ErbB	p. 53
5.2	<i>Network</i> levi	p. 54
5.3	Objeto Ordering Data	p. 56
5.4	Objeto Modularity Data	p. 57
5.5	Tela principal de execução do <i>dashboard</i>	p. 59
5.6	Tela da opção <i>Cluster plot</i>	p. 60
5.7	Área de visualização do gráfico gerado a partir dos <i>clusters</i>	p. 60
5.8	Tela inicial da opção do menu <i>Gene expression Level</i>	p. 61
5.9	Aba <i>Biomart and Suavization</i> da opção do menu <i>Gene Expression Level</i> . . .	p. 63
5.10	Aba <i>Plot Settings</i>	p. 64
5.11	Gráficos de nível de expressão gênica - agrupamento	p. 64
5.12	Gráficos de nível de expressão gênica - unitário	p. 65
5.13	Aba <i>Targets</i> do menu <i>Functional Enrichment</i>	p. 65
5.14	Aba <i>Ontology and Statistic</i> do menu <i>Functional Enrichment</i>	p. 66
5.15	Aba <i>Comparison Results</i> do menu <i>Functional Enrichment</i>	p. 67
5.16	Aba <i>Graph</i> do menu <i>Functional Enrichment</i>	p. 67
5.17	Exemplo de visualização do levi dentro do <i>dashboard</i>	p. 68

5.18	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Cruz.	p. 69
5.19	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo Anel.	p. 69
5.20	Gráfico de perfil de modularidade utilizando as janelas 101, 201, 301, 401 e 501 para o modelo X.	p. 70
5.21	Gráfico de Custo Energético em relação a quantidade de passos de Monte Carlo.	p. 70
5.22	Gráfico de modularidade por posição relativa das proteínas.	p. 72
5.23	Seleção de clusters para criação de sub-redes.	p. 72
5.24	Sub-redes criadas a partir da seleção dos módulos.	p. 74
5.25	Análise da atividade transcricional relativa da série GSE10072	p. 76
5.26	Análise da atividade transcricional relativa da série GSE10072	p. 76
5.27	Diagrama de Veen para a as comparações da série GSE10072	p. 77
5.28	Aplicação do levi para visualização da sobreposição do cluster AC em dados de expressão da série GSE10072.	p. 78
5.29	Análise da atividade transcricional relativa da série GSE10072	p. 79
5.30	Análise da atividade transcricional relativa da série GSE10072	p. 79
5.31	Diagrama de Veen para a as comparações da série GSE81089	p. 80
5.32	Aplicação do levi para visualização da sobreposição do cluster AC em dados de expressão da série GSE81089.	p. 81
5.33	Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória. . .	p. 88
5.34	Transcriptograma da comparação 2 (ex-fumantes com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 89
5.35	Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE10072 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 90

5.36	Transcriptograma da comparação 1 (fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória. .	p. 92
5.37	Transcriptograma da comparação 2 (ex-fumantes com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 93
5.38	Transcriptograma da comparação 3 (indivíduos que nunca fumaram com câncer) para a série GSE81089 com a apresentação de algumas ontologias selecionadas de forma aleatória.	p. 94

Lista de Tabelas

4.1	Detalhes dos dados de expressão utilizados na análise	p. 30
5.1	Número de <i>downloads</i> por ano e mês do pacote <i>levi</i> no Bioconductor.	p. 55
5.2	Posições relativas das proteínas para formação dos <i>clusters</i> selecionados.	p. 73
5.3	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE10072.	p. 82
5.4	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE10072.	p. 83
5.5	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE10072.	p. 84
5.6	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer fumantes da série GSE81089.	p. 85
5.7	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer ex-fumantes da série GSE81089.	p. 86
5.8	Amostra de processos biológicos existentes no <i>cluster</i> 1 com p-valor menor que 1×10^{-5} de indivíduos com câncer que nunca fumaram da série GSE81089.	p. 87

Sumário

Resumo	p. iv
Abstract	p. v
1 Introdução	p. 1
2 Revisão da Literatura	p. 3
2.1 Big Data	p. 3
2.2 Biologia de sistemas	p. 5
2.3 Tecnologias para obtenção e análise dos dados de expressão gênica	p. 8
2.3.1 Microarranjo	p. 8
2.4 Tecnologias de Microarranjo	p. 8
2.4.1 Affymetrix	p. 8
2.4.2 Roche	p. 9
2.4.3 Agilent	p. 10
2.4.4 RNA-seq	p. 11
2.4.5 Sanger (primeira-geração)	p. 12
2.4.6 Sequenciamento de nova geração	p. 14
2.4.7 Single Cell	p. 17
2.5 Normalização dos dados	p. 20
2.6 Dados transcriptômicos	p. 20
2.6.1 <i>Sequence Read Archive</i> (SRA)	p. 20
2.6.2 <i>Gene Expression Omnibus</i> (GEO)	p. 20
2.7 Redes	p. 21

2.7.1	<i>Search Tool for the Retrieval of Interacting Genes (STRING)</i>	p. 21
2.7.2	STITCH	p. 22
2.8	Redes Biológicas	p. 22
2.8.1	Conectividade	p. 24
2.8.2	Coeficiente de clusterização para redes não direcionadas	p. 24
2.8.3	Modelos de Redes	p. 24
2.8.4	Rede Escala Livre	p. 25
2.8.5	Rede Hierárquica	p. 26
2.8.6	Justificativa	p. 27
3	Objetivos	p. 29
3.1	Objetivos específicos	p. 29
4	Material e Métodos	p. 30
4.1	<i>Workflow</i>	p. 30
4.2	levi	p. 30
4.2.1	R	p. 32
4.2.2	C++	p. 33
4.2.3	CSS	p. 33
4.2.4	Equações	p. 33
4.3	<i>transcriptograma</i>	p. 37
4.4	Método de ordenamento de redes em uma dimensão	p. 37
4.4.1	Modelo Cruz	p. 40
4.4.2	Modelo X	p. 43
4.4.3	Modelo Anel	p. 44
4.5	Modularidade	p. 45
4.6	Análise de Expressão Gênica	p. 47

4.7	Enriquecimento Funcional	p. 48
4.8	Desenvolvimento da ferramenta <i>dashboard</i>	p. 48
4.8.1	Método transcriptograma	p. 48
5	Resultados e Discussão	p. 50
5.1	levi	p. 50
5.1.1	Implementação	p. 50
5.1.2	Instalação	p. 51
5.1.3	Carregando os arquivos	p. 52
5.1.4	Execução do levi	p. 52
5.1.5	Bioconductor	p. 55
5.2	transcriptograma	p. 55
5.3	Ferramenta de <i>dashboard</i>	p. 57
5.4	Aplicação do <i>workflow</i> em dados de <i>Homo sapiens</i>	p. 67
6	Conclusões	p. 96
	Referências Bibliográficas	p. 97

1 *Introdução*

A tecnologia da informação está cada vez mais presente em todas as áreas e aspectos de nossas vidas. Segundo (ALHARTHI, 2017), toda a atividade humana gera dados. No nosso dia a dia geramos dados nas interações que realizamos em redes sociais, com os equipamentos de GPS de celulares, ao acessar a redes *wi-fi* ou mesmo enviar mensagens em correios eletrônicos. Na área biológica, dados podem ser gerados de diferentes fontes como, por exemplo, os provenientes da interação entre proteínas ou os resultados obtidos através de tecnologias para o sequenciamento dos dados de expressão.

A integração e análise de dados computacionais obtidos de diferentes fontes é chamado de *big data* (SAGIROGLU; SINANC, 2013). Com o *big data* podemos realizar análises mais robustas incluindo dados que podem complementar resultados oriundos das mais diversas tecnologias.

Existem várias características do *big data* tal como os Vs que estão relacionados a atributos importantes que devemos levar em consideração para a correta análise dos diferentes conjuntos (variedade de dados) e a grande quantidade de dados existentes (volume). Para observarmos o nível de complexidade dos conjuntos, em uma análise de *big data* podemos exemplificar que a variedade dos dados podem ser estruturados, semi-estruturados ou não estruturados, um exemplo de dados não estruturados são as lojas virtuais que podem vender itens com diferentes características como tamanho, dimensão, volume, quantidade, cores, funcionalidades, entre outros. Em relação a característica do volume de dados podem ser medidos em terabytes, pentabytes e pode chegar até a *yottabytes* (RISTEVSKI; CHEN, 2018).

Para armazenar essa grande quantidade de dados de diferentes tipos existem repositórios disponíveis na *internet* como o IBGE com dados do senso brasileiro, o DATASUS com conjuntos de dados sobre o SUS e o Kaggle que possui dados desde filmes do cinema até desastres ambientais. Para dados obtidos através de diversas tecnologias ou experimentos biológicos podemos citar alguns exemplos como o *Gene Expression Omnibus*, *ArrayExpress* e o *The Cancer Genome Atlas*.

Para que possamos estudar a possibilidade de transferir as análises realizadas utilizando as técnicas de *big data* para a área biológica, podemos utilizar a biologia de sistemas. Segundo (SONIS; VILLA, 2019), a biologia de sistemas permite a integração entre áreas como a biologia, tecnologia e computação. As perguntas biológicas geram a necessidade de novas tecnologias, a aplicação das novas tecnologias tem como resultados novos dados e a computação, através de novos programas, realizam as análises desses dados.

Entre os grandes desafios da análise desses grandes volumes de dados estão as seleções dos conjuntos, integração e análises dos dados (KADADI et al., 2014). Por trabalhar com um grande volumes de dados que podem ter sido gerados de diferentes tecnologias, a etapa de análise de dados de *big data* apresenta uma grande dificuldade para os profissionais de ambas as áreas biológicas e/ou computacionais. Uma das principais etapas da análise de dados são as criações de visualizações através da filtragem e processamentos dos dados.

Apesar de tabelas possuírem geralmente um grau de detalhamento maior em relação a leitura dos dados, as imagens e gráficos facilitam a visualização e entendimento dos resultados por todos os usuários. Em 2017 Czauderna e Schreiber citam em seu trabalho a importância da visualização para apresentar e explorar grandes quantidades de dados biológicos.

Com a quantidade crescente de dados experimentais gerados e disponibilizados em repositórios na internet, faz-se necessário o acesso a métodos que permitam uma melhor visualização das potenciais relações entre estes registros.

6 *Conclusões*

A utilização do *dashboard* com as funções de ordenamento e modularidade permitiram a construção de sub-redes e visualização.

A criação de uma interface gráfica de usuário (GUI) integrando todas as ferramentas permitiu através de caixas de diálogos orientar o usuário sobre as etapas do *workflow* e possíveis mensagens de erro, além de ser mais amigável pela utilização de caixas de seleções para carregar os arquivos necessários e botões para ajustes nos parâmetros de execução dos processos. Foram minimizados os arquivos de entrada para a execução de todos os processos, sendo necessário somente o arquivo de interação dos nodos da rede biológica e os arquivos de expressão de microarranjo ou RNA-seq. As etapas de seleção de quantidade de comparações, controles e testes foram otimizadas para serem executadas diretamente do *dashboard*. Como arquivos de saída é possível salvar em variáveis da linguagem de programação R ou mesmo em arquivos de ambientes externos (.RData) os resultados de todas as etapas para uma análise mais analítica.

Ao realizarmos as análises com dados de expressão oriundos das séries GSE10072 (microarranjo) e GSE81089 (RNA-seq) obtivemos similaridades entre os resultados das comparações e conseguimos aumentar a acurácia em relação aos resultados publicados desses dados. Também foi possível analisar todo o organismo de uma forma mais ampla, sem descarte de genes, por utilizar como base a rede proteica como referência.

Referências Bibliográficas

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002.

ALHARTHI, A. *Addressing barriers to big data*. Amsterdam: Elsevier, 2017. 285-292 p.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, v. 25, n. 1, p. 25–29, May 2000. Disponível em: <<http://dx.doi.org/10.1038/75556>>.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, Nature Publishing Group, v. 5, n. 2, p. 101–113, 2004.

BLAZOTTI, A. A. *Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global*. Dissertação (Mestrado) — Instituto de Biociências de Botucatu, Universidade Estadual Paulista "Julio de Mesquita Filho", 2016.

BROWN, N. et al. Big data in drug discovery. *Progress in medicinal chemistry*, v. 57, p. 277–356, 2018. ISSN 0079-6468.

BRUGGEMAN, F. J.; WESTERHOFF, H. V. The nature of systems biology. *Trends in microbiology*, v. 15, p. 45–50, jan. 2007. ISSN 0966-842X.

CAMILO, E. et al. Galant: a cytoscape plugin for visualizing data as functional landscapes projected onto biological networks. *Bioinformatics*, Oxford University Press, v. 29, n. 19, p. 2505–2506, 2013.

CARBONE, F.; MONTECUCCO, F. Big data and data sharing: opportunities for the urgent challenges in cardiovascular disease. 2019. ISSN 0014-2972.

CASTRO, M. A. et al. Viacomplex: software for landscape analysis of gene expression networks in genomic context. *Bioinformatics*, Oxford University Press, v. 25, n. 11, p. 1468–1469, 2009.

CHARITOU, T.; BRYAN, K.; LYNN, D. J. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, BioMed Central, v. 48, n. 1, p. 27, 2016.

COSTA, F. F. Big data in biomedicine. v. 19, p. 433–440, 2014. ISSN 1359-6446.

CZAUDEANA, T.; SCHREIBER, F. Information visualization for biological data. *Methods in molecular biology (Clifton, N.J.)*, v. 1526, p. 403–415, 2017. ISSN 1940-6029.

DALMA-WEISZHAUSZ, D. D. et al. The affymetrix genechip platform: an overview. *Methods in enzymology*, v. 410, p. 3–28, 2006. ISSN 0076-6879.

- DEITEL, H. M.; DEITEL, P. J. *C++ como programar*. [S.l.]: Prentice Hall, 1999.
- DUDOIT, S. et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, JSTOR, p. 111–139, 2002.
- EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, Oxford Univ Press, v. 30, n. 1, p. 207–210, 2002.
- ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, v. 5, n. 1, p. 17–60, 1960.
- FABREGAT, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, v. 46, n. D1, p. D649–D655, 2018.
- FRASER, H. B. Modularity and evolutionary constraint on proteins. *Nat Genet*, London, UK, v. 37, n. 4, p. 351–352, Apr 2005. Disponível em: <<http://dx.doi.org/10.1038/ng1530>>.
- GOLDMANN, T. et al. Pd-l1 amplification is associated with an immune cell rich phenotype in squamous cell cancer of the lung. *Cancer immunology, immunotherapy : CII*, v. 70, p. 2577–2587, set. 2021. ISSN 1432-0851.
- GONZALEZ-ANGULO, A. M.; HENNESSY, B. T.; MILLS, G. B. Future of personalized medicine in oncology: a systems biology approach. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 28, n. 16, p. 2777–2783, 2010.
- GOODMAN, D. *Dynamic HTML: The Definitive Reference: A Comprehensive Resource for HTML, CSS, DOM & JavaScript*. [S.l.]: "O'Reilly Media, Inc.", 2002.
- HAN, Y. et al. Advanced applications of RNA sequencing and challenges. *Bioinformatics and biology insights*, v. 9, p. 29–46, 2015. ISSN 1177-9322.
- HEATHER, J. M.; CHAIN, B. The sequence of sequence: The history of sequencing DNA. *Genomics*, v. 107, n. 1, p. 1–8, January 2016.
- HRDLICKOVA, R.; TOLOUE, M.; TIAN, B. RNA-seq methods for transcriptome analysis. *Wiley interdisciplinary reviews. RNA*, v. 8, jan. 2017. ISSN 1757-7012.
- HUANG, Y.-W.; HUANG, T. H.-M.; WANG, L.-S. Profiling DNA methylomes from microarray to genome-scale sequencing. *Technology in cancer research & treatment*, v. 9, p. 139–147, abr. 2010. ISSN 1533-0338.
- HUBER, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, Nature Publishing Group, v. 12, n. 2, p. 115, 2015.
- HUGHES, T. R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature biotechnology*, v. 19, p. 342–347, abr. 2001. ISSN 1087-0156.
- HWANG, B.; LEE, J. H.; BANG, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, Nature Publishing Group, v. 50, n. 8, p. 1–14, 2018.

- IDEKER, T.; GALITSKI, T.; HOOD, L. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 2, n. 1, p. 343–372, 2001.
- JALALI, S. et al. Systematic transcriptome wide analysis of lncrna-mirna interactions. *PloS one*, Public Library of Science, v. 8, n. 2, p. e53823, 2013.
- KADADI, A. et al. Challenges of data integration and interoperability in big data. In: . Washington, DC, USA: IEEE, 2014. p. 38–40. ISBN 978-1-4799-5666-1.
- KANEHISA, M. et al. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, v. 40, n. Database issue, p. D109–D114, Jan 2012. Disponível em: <<http://dx.doi.org/10.1093/nar/gkr988>>.
- KHOURY, M. J.; IOANNIDIS, J. P. A. Big data meets public health. v. 346, p. 1054–1055, 2014. ISSN 0036-8075.
- KUHN, M. et al. Stitch 4: integration of protein–chemical interactions with user data. *Nucleic acids research*, Oxford Univ Press, p. gkt1207, 2013.
- LANDI, M. T. et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*, Public Library of Science, v. 3, n. 2, p. e1651, 2008.
- LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. *Nucleic acids research*, Oxford Univ Press, p. gkq1019, 2010.
- LIU, J. et al. A prognostic 14-gene expression signature for lung adenocarcinoma: A study based on tcga data mining. *Oxidative medicine and cellular longevity*, v. 2020, p. 8847226, 2020. ISSN 1942-0994.
- LIU, L. et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, v. 2012, p. 251364, 2012. ISSN 1110-7251.
- LIU, Y. et al. Systematic analysis of mrna expression profiles in nscic cell lines to screen metastasis-related genes. *Molecular medicine reports*, v. 14, p. 5093–5103, dez. 2016. ISSN 1791-3004.
- LUN, A. T. L.; MCCARTHY, D. J.; MARIONI, J. C. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, v. 5, p. 2122, 2016. ISSN 2046-1402.
- METZKER, M. L. Sequencing technologies - the next generation. *Nature reviews genetics*, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010.
- MILLER, M. B.; TANG, Y.-W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, Am Soc Microbiol, v. 22, n. 4, p. 611–633, 2009.
- MOLAN, A. L.; RYBARCZYK-FILHO, J. L. Desenvolvimento e compara  o de algoritmos para a organiza  o hier rquica de redes. *Anais do X Congresso de F sica Aplicada   Medicina*, v. 1, n. 117-121, 2014.

MOREIRA, L. M. *Ciencias genômicas : fundamentos e aplicacoes*. Sociedade Brasileira de Genetica, 2015. Disponível em: <<https://drive.google.com/file/d-/0Bwb0CldyHPwV3IxRGdyYklIdGc/view>>.

NELSON, D. L.; COX, M. M. *Lehninger Principles of Biochemistry, Fourth Edition*. Fourth edition. New York, USA: [s.n.], 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>.

RISTEVSKI, B.; CHEN, M. Big data analytics in medicine and healthcare. *Journal of integrative bioinformatics*, De Gruyter, v. 15, n. 3, 2018.

RYBARCZYK-FILHO, J. L. et al. Towards a genome-wide transcriptogram: the *saccharomyces cerevisiae* case. *Nucleic acids research*, Oxford Univ Press, v. 39, n. 8, p. 3005–3016, 2011.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: . San Diego, CA, USA: IEEE, 2013. p. 42–47. ISBN 978-1-4673-6402-7.

SANCHEZ-PINTO, L. N.; LUO, Y.; CHURPEK, M. M. Big data and data science in critical care. *Chest*, v. 154, p. 1239–1248, nov. 2018. ISSN 1931-3543.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 12, p. 5463–5467, 1977.

SHAHZAD, K.; LOOR, J. J. Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. *Current genomics*, v. 13, p. 379–394, ago. 2012. ISSN 1389-2029.

SONIS, S.; VILLA, A. *Translational Systems Medicine and Oral Disease*. [S.l.]: Academic Press, 2019.

STRING. *STRING*. 2017. <http://string-db.org/>. [Online; accessed 17-Dec-2019].

SU, H. et al. Identification of potential biomarkers of lung adenocarcinoma brain metastases via microarray analysis of cDNA expression profiles. *Oncology letters*, v. 17, p. 2228–2236, fev. 2019. ISSN 1792-1074.

SZCZEŚNIAK, M. W.; MAKALOWSKA, I. lncRNA-rna interactions across the human transcriptome. *PloS one*, Public Library of Science, v. 11, n. 3, p. e0150353, 2016.

SZKLARCZYK, D. et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, Oxford Univ Press, p. gku1003, 2014.

TANG, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nature methods*, v. 6, p. 377–382, maio 2009. ISSN 1548-7105.

TREVINO, V.; FALCIANI, F.; BARRERA-SALDANA, H. A. Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular medicine (Cambridge, Mass.)*, v. 13, p. 527–541, 2007. ISSN 1076-1551.

TSENG, G. C.; GHOSH, D.; FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, v. 40, p. 3785–3799, maio 2012. ISSN 1362-4962.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

VINOGRADOV, A. E. Modularity of cellular networks shows general center-periphery polarization. *Bioinformatics*, Oxford, UK, v. 24, n. 24, p. 2814–2817, Dec 2008. Disponível em: <<http://dx.doi.org/10.1093/bioinformatics/btn555>>.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, v. 10, p. 57–63, jan. 2009. ISSN 1471-0064.

WINNICK, E. Dna sequencing industry sets its sights on the future: new technologies may duke it out with upgrades of the current method. *The Scientist*, Scientist Inc., v. 18, n. 18, p. 44–48, 2004.