RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta dissertação será disponibilizado somente a partir de 17/02/2019.

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

Alex Augusto Biazotti

Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global

Botucatu
Fevereiro/2017

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

Alex Augusto Biazotti

Desenvolvimento de Ferramenta Computacional para integração de transcriptomas e redes biológicas: medidas de desempenho global

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

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho
Co-Orientadora: Prof.ª Dr.ª Agnes Alessandra Sekijima Takeda

Botucatu
Fevereiro/2017

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

Biazotti, Alex Augusto.
Desenvolvimento de ferramenta computacional para integração de transcriptomas e redes biológicas : medidas de desempenho global / Alex Augusto Biazotti. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Coorientador: Agnes Alessandra Sekijima Takeda
Capes: 10302026

1. Transcriptoma. 2. Ontologia. 3. Análise de microarranjo. 4. Processamento eletrônico de dados.

Palavras-chave: Integração de dados; Microarranjo; Ontologia; Rede proteica; Transcriptograma.

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 e Dra. Agnes Alessandra Sekijima Takeda pela excelente orientação, apoio e paciência;
- Ao meu pai José e a minha mãe Cleide pelo apoio e incentivo durante todo esse tempo;
- Aos meus amigos André Luiz Molan, Carlos Alberto de Oliveira Biagi Junior e José Rafael Pilan pela amizade e ajuda na confecção deste trabalho;
- Aos demais professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho;
- À Deus pela vida.

Resumo

A cada dia surgem novas tecnologias que possibilitam o estudo em larga escala dos RNAs transcritos por um organismo em condições específicas, com isso fornecendo uma grande quantidade de informações. No entanto as metodologias tradicionais não são capazes de análisar de forma eficiente esses dados por utilizar *cut-offs* pré-definidos, eliminando assim uma grande quantidade de genes não considerados diferencialmente expresso, e por consequência reduzindo a precisão e a acurácia do estudo. Esse trabalho propõe o aperfeiçoamento da metodologia do transcriptograma (modelo cruz), desenvolvido por Rybarczyk-Filho *et al.*, que realiza uma análise de forma global de um organismo, utilizando por sua vez redes proteícas e processos biológicos. Dentre as modificações realizadas estão: mudança no algoritmo de ordenamento para a redução do tempo de processamento da rede, adição de dois novos modelos "X" e "Anel", a automação dos processos de análise de dados de expressão gênica, enriquecimento funcional e da compilação de todas as informações em um gráfico. Para testar o aperfeiçoamento foram utilizadas duas séries de dados de expressão gênica, a GSE10072 e a GSE19804, referentes a amostras de câncer de pulmão. O modelo "Anel" apresentou a melhor redução do custo energético de uma matriz, aproximadamente 93%. Para a modularidade, o modelo "Anel" também teve o melhor desempenho. A automação dos processos de enriquecimento funcional, da análise dos dados de expressão e da compilação de todos os dados em forma gráfica diminui o tempo gasto para a aquisição e geração, além de aumentar a acurácia. Os resultados indicam que independentemente do hábito ou nacionalidade de um indivíduo, um mesmo tipo de câncer podem apresentar os mesmos conjuntos de processos biológicos alterados. A ferramenta não encontrou os mesmos processos biológicos indicados pelos *software* PAGE e GAGE, porém ele retornou processos mães ou filhos dos mesmos. A utilização desta ferramenta pode ser uma nova alternativa comparado aos demais métodos, devido a utilização de diversas informações adicionais ao conjunto de expressão gênica a ser analisado.

Abstract

Every day, new technologies are emerging that make it possible the large-scale study of RNAs transcribed by an organism under specific conditions, providing a huge amount of information. However, the traditional methodologies are not able to efficiently analyze these data due the use of pre-defined cut-offs, thus eliminating a large number of genes not considered differentially expressed, and consequently reducing precision andaccuracy of the study. This work proposes the improvement of the methodology of the Transcriptogram (model "Cross"), developed by Rybarczyk-Filho *et al.*, which performs an overall analysis of an organism, using protein networks and biological processes. Among the modifications made are: Modification in ordering algorithm to reduce the network processing time, addition of the two new "X" and "Ring" models, the automation of the processes of gene expression data analysis, functional enrichment and the compilation of all information in a graphic. To test the improvements, two sets of gene expression data were used, GSE10072 and GSE19804, corresponding to samples of lung cancer. The "Ring" model showed the best matrix energy cost reduction, approximately 93%. For modularity, the "Ring" model also had the best performance. The automation of functional enrichment processes, the analysis of expression data and the compilation of all data in a graphic form reduces the time spent for acquisition and generation, increasing the accuracy. The results indicate that regardless of habit of an individual, the same type of cancer may present the same sets of altered biological processes. The tool did not find the same biological processes indicated by the software PAGE and GAGE, but it returned their ancestor or child processes. The use of this tool may be a new alternative to the other methods, due the use of additional information to the set of gene expression to be analyzed.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Microarranjo

No final do século XX, os pesquisadores tinham dificuldades de medir a expressão de vários genes de um organismo ao mesmo tempo, nesta época era possível medir apenas a expressão de poucos genes por vez. Mas com o passar dos anos, novas tecnologias foram desenvolvidas, uma dessas tecnologias foi a criação de um *chip* contendo várias sequências de nucleotídeos (cDNA ou oligonucleotídeo) denominado sonda. Com isso esse *chip* tornou-se uma ferramenta padrão para muitos laboratórios de pesquisa genômica(TSENG; GHOSH; FEINGOLD, 2012).

Um *chip* de microarranjo é composto por *spots/beads* e sondas. Os *spots/beads* são divisões no *chip* de microarranjo com identificadores que contém apenas parte de uma sequência com diversas cópias da mesma, denominada sonda, cada sonda é composta de 20-60 oligonucleotídeos (Figura 1.1), e ela é capaz de hibridizar com um fragmento de gene (Figura 1.2). A quantidade de *spot/bead* é diferente para cada organismo.

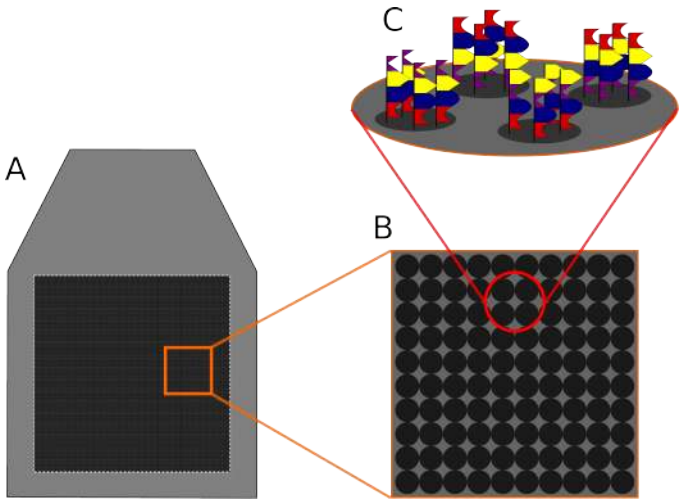

Figura 1.1: Representação dos componentes de um chip de microarranjo. (A) *chip* de microarranjo. (B) Pontos mais escuros, geralmente pretos, presentes no *chip* denominados *spots* ou *bead*. (C) Sondas presentes nos *spots*, onde cada sonda apresenta de 20-60 oligonucleotídeos.

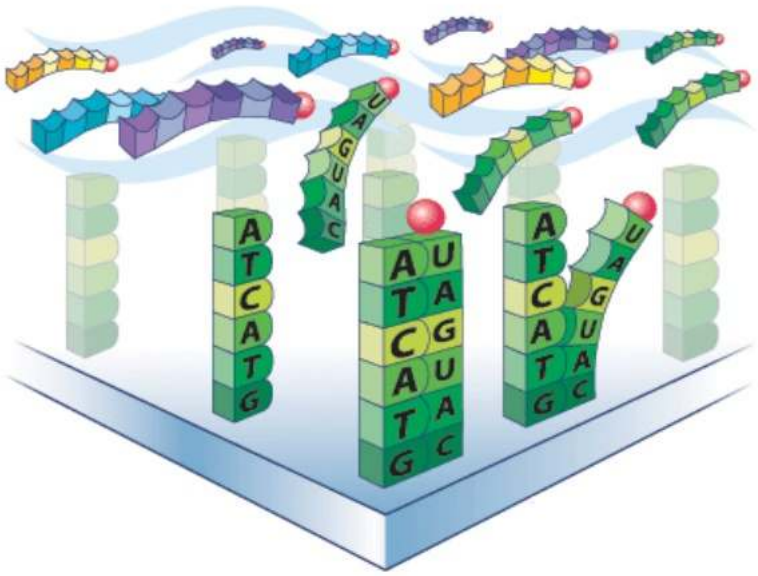

Figura 1.2: Hibridização da sonda do microarranjo com o fragmento do gene, representado pela sequência de nucleotídos com uma esfera vermelha. Os nucleotídeos presentes na sonda do chip irão combinar com os nucleotídeos presentes no fragmento dos genes. Adaptado de `www.essex.ac.uk/staff/langdon/genechip/`

A tecnologia mais utilizada realiza a hibridização de duas amostras (referência e teste). Com as sequências presentes no *chip* (LEUNG; CAVALIERI, 2003) e através de cálculos matemáticos e estatísticos obtêm-se os dados de expressão de cada sonda (XIE; PAN; KHODURSKY, 2005). Para a obtenção das informações de expressão é necessário duas amostras(referência e teste). Em seguida é realizada a extração dos mRNAs (RNA mensageiros) das amostras, então Aplica-se a enzima transcriptase reversa para obter o cDNAs, durante a obtenção dos cDNAs são utilizados nucleotídeos com os marcadores fluorescentes, o marcador vermelho para os mRNAs referentes a amostra do caso e verdes para a amostra referência. Com a obtenção dos cDNAs com os marcadores fluorescentes realiza-se a hibridização dos cDNAs com o *chip* de microarranjo, deixando eles agirem por algumas horas. Após a hibridização o chip é lavado para a remoção de cDNAs não hibridizados e colocado em um *scanner* que irá emitir um laser sobre o *chip*, essa emissão realizada duas vezes, onde uma vez irá emitir na frequência para captar a tonalidade vermelha e depois na frequência para captar a tonalidade verde. Depois da captação, as mesmas são mescladas através de um algoritmo estatístico, onde apresenta novas colorações como tons que variam de amarelo até laranja, essa nova variação a representação que houve a hibridização tanto da amostra teste quanto da amostra referência naquele *spot/bend*. Entretanto ele pode apresentar a coloração preta que é referente a não expressão de nenhuma das amostras (Figura 1.3).

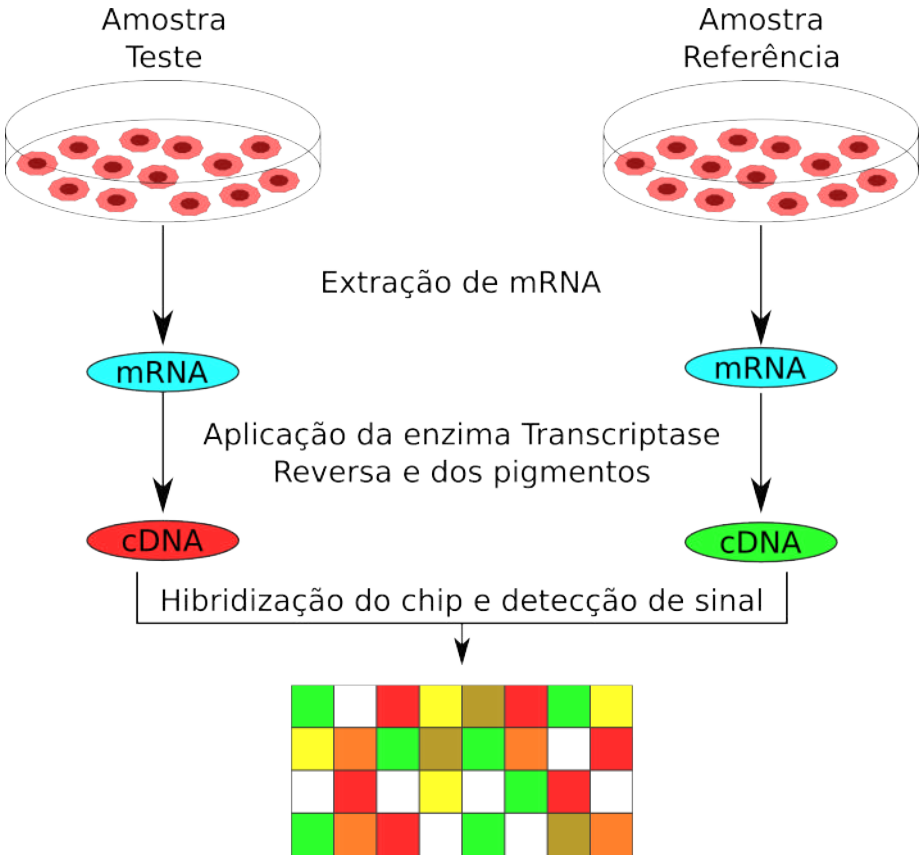

Figura 1.3: Representação referente a extração da informação da expressão dos genes, extração do mRNA das amostras e aplicação da enzima transcriptase reversa juntamente com nucleotídeos com marcadores fluorescentes verde e vermelho. Por fim são combinados e aplicados no chip, deixando hibridizar por algum tempo e inseridos em um sistema computadorizado para a extração da luminosidade dos genes hibridizados.

Com a aquisição dos níveis de expressão registrados pelo *chip*, através da frequência luminosa, é necessária a extração dos dados através de uma técnica de normalização para que seja possível a manipulação/estudo da expressão dos genes. Existem diversas técnicas de normalização, mas as mais utilizadas são a *Affymetrix Microarray Suite 5.0* (MAS5), *Robust Multi-array Analysis* (RMA) e *Robust Multi-array Analysis with correction for GC content* (GCRMA). Por meio destas técnicas torna-se possível o estudo da expressão dos genes por meio de cálculos como *fold-change*, expressão média, teste-t, p-valor para cada gene (DALMAN et al., 2012), e com o resultados desses cálculos, os pesquisadores são capazes de aplicar alguns critérios nos valores para verficação de quais genes estão superexpressos ou subexpressos.

5 *Conclusões*

A alteração realizada no algoritmo de ordenamento, inserindo os arquivos de verificação do processamento dos dados e a alteração da matriz adjacente, teve uma boa performance em comparação aos outros ordenamentos, sendo esse capaz de reduzir de forma drástica o tempo necessário para o ordenamento, quando comparado com os ordenamentos desenvolvidos por (RYBARCZYK-FILHO et al., 2011) e (MOLAN; RYBARCZYK-FILHO, 2014), sem haver perda do poder de redução do custo energético. Ao analisar todos modelos de análise de vizinhança, verificou-se que o modelo "Anel" apresentou os melhores resultados tanto para a redução do custo energético quanto para a clusterização e aproximação da diagonal principal.

Através da automação do processo de enriquecimento funcional, que anteriormente era realizado de forma manual e verificando um-a-um em determinados sites, fez com que reduzisse o tempo necessário para encontrar os processos biológicos referentes as proteínas com um p-valor igual ou inferior a 1×10^{-5} dentro de cada módulo. Além disso aumentamos a acurácia dos resultados por considerar apenas os processos biológicos que apresentavam no mínimo 60% das proteínas dos processos presentes no módulo.

Ao aplicar a metodologia em amostras de câncer em diferentes indivíduos, foi possível verificar uma certa semelhança existente entre esses indivíduos, sendo a superexpressão e subexpressão muito próximos, levando em conta que o nível de expressão das proteínas ainda são diferentes. Além disso muitos processos biológicos são semelhantes nos 4 grupos análisados.

Apesar de alguns processos ainda necessitarem de manipulação manual do usuário, como a seleção dos módulos e a determinação dos parametros, as análises finais apresentam resultados com excelente qualidade. A metodologia apresenta um grande diferencial em relação as outras metodologias de análise de expressão gênica, pois é realizada utilizando uma rede proteíca que permite analisar o organismo de uma forma global. Em comparação aos outros que realizam de forma mais pontual, ou seja, apenas nos genes, além de descartarem genes que não são considerados diferencialmente expressos apresentam uma grande quantidade falsos positivos.

Referências Bibliográficas

BARNES, M. Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms. *Nucleic Acids Research*, Oxford University Press (OUP), v. 33, n. 18, p. 5914?5923, Oct 2005. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gki890>.

BARRETT, T. et al. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Res*, v. 41, n. Database issue, p. D991–D995, Jan 2013. Disponível em: <http://dx.doi.org/10.1093/nar/gks1193>.

CONSORTIUM, G. O. Gene ontology consortium: going forward. *Nucleic Acids Res*, v. 43, n. Database issue, p. D1049–D1056, Jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar-/gku1179>.

CUI, X.; CHURCHILL, G. A. Statistical tests for differential expression in cdna microarray experiments. *Genome biology*, BioMed Central, v. 4, n. 4, p. 1, 2003.

DALMAN, M. R. et al. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, v. 13 Suppl 2, p. S11, Mar 2012. Disponível em: <http://dx.doi.org/10.1186/1471-2105-13-S2-S11>.

DU, P.; KIBBE, W. A.; LIN, S. M. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, Oxford University Press (OUP), v. 24, n. 13, p. 1547?1548, May 2008. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btn224>.

DUNNING, M. J. et al. beadarray: R classes and methods for illumina bead-based data. *Bioinformatics*, Oxford University Press (OUP), v. 23, n. 16, p. 2183?2184, Jun 2007. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btm311>.

FAN, J. et al. [3] illumina universal bead arrays. *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, Elsevier BV, p. 57?73, 2006. ISSN 0076-6879. Disponível em: <http://dx.doi.org/10.1016/S0076-6879(06)10003-8>.

GAUTIER, L. et al. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, Oxford University Press (OUP), v. 20, n. 3, p. 307?315, Feb 2004. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/btg405>.

GERSHON, D. Microarray technologyan array of opportunities. *Nature*, Springer Nature, v. 416, n. 6883, p. 885?891, Apr 2002. ISSN 0028-0836. Disponível em: <http://dx.doi.org-/10.1038/416885a>.

GHARAIBEH, R. Z.; FODOR, A. A.; GIBAS, C. J. Background correction using dinucleotide affinities improves the performance of gcrma. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 1, 2008.

GUSNANTO, A.; CALZA, S.; PAWITAN, Y. Identification of differentially expressed genes and false discovery rate in microarray studies. *Current opinion in lipidology*, LWW, v. 18, n. 2, p. 187–193, 2007.

IRIZARRY, R. A. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, Oxford University Press (OUP), v. 31, n. 4, p. 15e?15, Feb 2003. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gng015>.

IRIZARRY, R. A. et al. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, v. 31, n. 4, p. e15, Feb 2003.

IRIZARRY, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, Biometrika Trust, v. 4, n. 2, p. 249–264, 2003.

KIM, C. C. et al. Improved analytical methods for microarray-based genome-composition analysis. *Genome biology*, BioMed Central, v. 3, n. 11, p. 1, 2002.

KIM, S.-Y.; VOLSKY, D. J. Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, BioMed Central Ltd, v. 6, n. 1, p. 144, 2005.

KUENTZER, F. A. *Otimização e análise de algoritmos de ordenamento de redes proteicas*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2014.

LANDI, M. T. et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, v. 3, n. 2, p. e1651, Feb 2008. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0001651>.

LEUNG, Y. F.; CAVALIERI, D. Fundamentals of cdna microarray data analysis. *Trends in Genetics*, Elsevier BV, v. 19, n. 11, p. 649?659, Nov 2003. ISSN 0168-9525. Disponível em: <http://dx.doi.org/10.1016/j.tig.2003.09.015>.

LOCKHART, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, Springer Nature, v. 14, n. 13, p. 1675?1680, Dec 1996. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1296-1675>.

LU, T.-P. et al. Identification of a novel biomarker, sema5a, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*, v. 19, n. 10, p. 2590–2597, Oct 2010. Disponível em: <http://dx.doi.org/10.1158/1055-9965.EPI-10-0332>.

LUO, W. et al. Gage: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, v. 10, p. 161, 2009. Disponível em: <http://dx.doi.org/10.1186/1471-2105-10-161>.

MOLAN, A. L.; RYBARCZYK-FILHO, J. L. Desenvolvimento e comparaï¿½ï¿½o de algoritmos para a organizaï¿½ï¿½o hierï¿½rquica de redes. *Anais do X Congresso de Fï¿½sica Aplicada ï¿½ Medicina*, v. 1, n. 117-121, 2014.

NAEF, F.; MAGNASCO, M. O. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E*, APS, v. 68, n. 1, p. 011906, 2003.

PEPPER, S. D. et al. The utility of mas5 expression summary and detection call algorithms. *BMC bioinformatics*, BioMed Central, v. 8, n. 1, p. 1, 2007.

RYBARCZYK-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>.

STEEMERS, K. L. G. F. J. Illumina, inc. *Pharmacogenomics*, v. 6, n. 7, p. 777–782, Oct 2005.

SZKLARCZYK, D. et al. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, v. 43, n. Database issue, p. D447–D452, Jan 2015. Disponível em: <http://dx.doi.org/10.1093/nar/gku1003>.

TRIPATHI, S.; GLAZKO, G. V.; EMMERT-STREIB, F. Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. *Nucleic Acids Res*, v. 41, n. 7, p. e82, Apr 2013. Disponível em: <http://dx.doi.org/10.1093-/nar/gkt054>.

TSENG, G. C.; GHOSH, D.; FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, Oxford University Press (OUP), v. 40, n. 9, p. 3785?3799, Jan 2012. ISSN 1362-4962. Disponível em: <http://dx.doi.org/10.1093/nar/gkr1265>.

VERLI, H. et al. Bioinformática da biologia à flexibilidade molecular. *Porto Alegre, Brasil*, v. 1, 2014.

WU, Z. A review of statistical methods for preprocessing oligonucleotide microarrays. *Statistical methods in medical research*, SAGE Publications, v. 18, n. 6, p. 533–541, 2009.

XIE, Y.; PAN, W.; KHODURSKY, A. B. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, Oxford University Press (OUP), v. 21, n. 23, p. 4280?4288, Sep 2005. ISSN 1460-2059. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti685>.

YANG, Y. H.; THORNE, N. P. Normalization for two-color cdna microarray data. *Lecture Notes-Monograph Series*, JSTOR, p. 403–418, 2003.

ZAHURAK, M. et al. Pre-processing agilent microarray data. *BMC Bioinformatics*, Springer Nature, v. 8, n. 1, p. 142, 2007. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186-/1471-2105-8-142>.

ZIMMERMANN, K.; LESER, U. Analysis of affymetrix exon arrays. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik, 2010.