



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Campus de São José do Rio Preto

TIAGO TAMBONIS

**ANÁLISE DO MÉTODO SUVREL NA
EXPRESSÃO DIFERENCIAL A PARTIR DA
MATRIZ DE CONTAGENS GERADA COM
DADOS DE RNA-SEQ**

SÃO JOSÉ DO RIO PRETO - SÃO PAULO

2015

TIAGO TAMBONIS

**ANÁLISE DO MÉTODO SUVREL NA EXPRESSÃO
DIFERENCIAL A PARTIR DA MATRIZ DE CONTAGENS
GERADA COM DADOS DE RNA-SEQ**

Dissertação apresentada para obtenção do título de Mestre em Biofísica Molecular, área de Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto.

Orientador: Prof. Dr. Vitor B. Pereira Leite

São José do Rio Preto - São Paulo

2015

Tambonis, Tiago.

Análise do método Suvrel na expressão diferencial a partir da matriz de contagens gerada com dados de RNA-Seq / Tiago Tambonis. -- São José do Rio Preto, 2015

69 f. : il.

Orientador: Vitor B. Pereira Leite

Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Expressão gênica.
4. Seqüenciamento de nucleotídeo. 5. Pesquisa quantitativa. I. Leite, Vitor Barbanti Pereira. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas.
III. Título.

CDU – 575.113

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

TIAGO TAMBONIS

**ANÁLISE DO MÉTODO SUVREL NA EXPRESSÃO
DIFERENCIAL A PARTIR DA MATRIZ DE CONTAGENS
GERADA COM DADOS DE RNA-SEQ**

Dissertação apresentada para obtenção do título de Mestre em Biofísica Molecular, área de Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto.

BANCA EXAMINADORA

Prof. Dr. Vitor B. Pereira Leite
Professor Livre Docente
UNESP – São José do Rio Preto – SP
Orientador

Claudia Marcia Aparecida Carareto
Professora Titular
UNESP – São José do Rio Preto – SP

Francisco Pereira Lobo
Pesquisador
Embrapa Informática Agropecuária –
Campinas – SP

São José do Rio Preto, 19 de maio de 2015

Dedico este trabalho a toda minha família, principalmente a minha mãe Rosane, minha irmã Priscila, as minhas avós Yolanda e Mirtes e a minha namorada Bianca, que sempre estiveram ao meu lado nesta etapa da minha vida.

Agradecimentos

A realização desta dissertação só foi possível graças ao apoio e colaboração de muitas pessoas e algumas delas infelizmente eu não vou conseguir lembrar mas, eu sou grato pela ajuda.

Agradeço a minha família por todo apoio, compreensão e amor. Agradeço sobretudo as mulheres que marcaram a minha vida: minha mãe, Rosane, minha irmã Priscila, minhas avós Mirtes e Yolanda e minha namorada, Bianca. Agradeço ao meu grande amigo Vitor e sua família, responsáveis por apoio indispensável em determinadas etapas da minha vida. Agradeço ao meu amigo Luiz Aurélio, pela amizade e por ter me mostrado a filosofia budista de Nitiren Daishonin, mudando definitivamente os rumos da minha vida. Agradeço ao meu amigo, Rafael Franco, por ter me apoiado em pensar em deixar o curso de engenharia de telecomunicações.

Agradeço ao meu orientador Vitor Barbanti Pereira Leite pela oportunidade cedida e por me orientar da melhor maneira possível. Agradeço de maneira geral ao grupo de pesquisa que faço parte, por sempre me ajudar a solucionar minhas dúvidas. Agradeço aos meus amigos do Departamento de Física que convivo diariamente por construírem um ambiente de trabalho muito sadio, agradável e familiar: Vinícius, Vinícius (Goiás), Josimar, Ingrid, Fernanda, Gabi, Tiago, Marcelo, Mirian, Laís, Alenxandre (Jesus), Daniel, Fernando, Monique, Taísa, Gabriel (Bibi), Natalia, Paulo, Rafaela, Tabata, Carol, Bruno e Icaro.

Agradeço também a todos os funcionários do IBILCE/UNESP por fazerem desta instituição um ótimo lugar para estudar, ao CNPq pelo apoio financeiro, e aos docentes que se esforçaram para me guiar no caminho do conhecimento.

Agradeço ao meu Mestre Daisaku Ikeda por difundir o verdadeiro budismo de Nitiren Daishonin me proporcionando a oportunidade de conhecer essa filosofia, e a todos meus companheiros da BSGI por me apoiarem a todo momento, em especial os amigos da Comunidade Renascença.

Agradeço aos membros da banca do exame geral de qualificação, professora Fátima Pereira de Souza e o Fábio Rogério de Moraes, pela disponibilidade em ler o texto aqui apresentado e sugerir valiosas alterações para a melhoria deste texto e do estudo. Agradeço por fim, a Professora Claudia Márcia Aparecida Carareto e o Francisco Pereira Lobo, por terem aceitado participar da banca de defesa.

“Eu não me envergonho de corrigir meus erros e mudar minhas opiniões, porque não me envergonho de raciocinar e aprender.”

Alexandre Herculano (1810-1877)

Resumo

Estamos vivendo uma época onde os avanços das áreas ligadas a biologia são rotineiros, nos levando cada vez mais a nos habituar a experimentos com um grande número de variáveis. A tecnologia de sequenciamento de RNA (RNA-Seq) é parte deste quadro e as abordagens computacionais aplicadas neste âmbito não estão totalmente estabelecidas e necessitam de análises mais detalhadas. A partir da tabela de contagens, que sumariza cada biblioteca em uma condição experimental, propõe-se a utilização de um método variacional chamado de Suvrel, baseado na minimização de uma função custo que penaliza grandes distâncias entre elementos de mesma classe e favorece pequenas distâncias entre elementos de classes diferentes, para inferência de expressão diferencial. A aplicação do método foi realizada em uma tabela de contagens produzida após o sequenciamento, alinhamento e sumarização de 5 replicatas técnicas de RNA de referência humano juntamente com a mistura ERCC 1 e 5 replicatas técnicas de RNA de referência do cérebro humano juntamente com a mistura ERCC 2. Utilizando curvas ROC produzidas com os dados do projeto do MAQC-II, definindo os transcritos analisados pelo projeto com \log_2 do *fold-change* maior que um limiar que varia de 0,5 a 2,0 como os verdadeiros positivos e os restantes como verdadeiros negativos, é possível concluir que o método Suvrel tem maiores valores abaixo das curvas ROC na maior parte dos limiares. Utilizando curvas ROC produzidas com os dados do ERCC, geradas utilizando o *logs* das mudanças das proporções predefinidas das misturas ERCC 1 e 2 de 92 oligonucleotídeos, é possível concluir que o método Suvrel tem a maior área abaixo da curva ROC. Embora as áreas abaixo das curvas ROC sejam comparáveis às de outros pacotes (como por exemplo o edgeR), é importante ressaltar que elas foram produzidas usando um método que não faz nenhum tipo de suposição quanto a distribuição associada aos *reads*, utilizando uma normalização simples (contagens de um gene com média 0 e variância 1) tendo rápida execução.

Palavras-chave: RNA-Seq. Expressão diferencial. Método Suvrel. Análise.

Abstract

We are living in a time where advances in areas related to biology are routine, taking us to accustom to experiments with large number of variables. The RNA sequencing technology (RNA-Seq) is part of this framework and computational approaches applied in this context are not fully established and require more detailed analysis. Generally, in a experiment of analysis of differential expression, total RNA samples or messengers (mRNA) is extracted, purified, fragmented, sequenced, mapped, and finally counted, generating an count table that relates how many reads was aligned to a given gene in a experimental condition. From this stage, it is proposed to use a variational method, called Suvrel (Supervised Variational Relevance), based on the minimization of a cost function that penalizes large distances between the same class of elements and favors small distances between different classes of elements to make the inference of relevance of each gene. The application of the method was performed on count table produced after of sequencing, alignment and summarization of 5 technical replicates containing Strategene Universal Human Reference RNA (UHRR) (part of Sequencing Quality Control Consortium, SEQC) together with ERCC 1 mix, and 5 technical replicates containing Ambion's Human Brain Reference RNA (HBRR) (part of SEQC also) together with the ERCC 2 mix. Using the ROC (Receiver Operating characteristic) curves generating from data of MAC-II project, setting the transcripts with log of fold-change greater than a cutoff (from 0.5 to 2.0) as true positive and the others as true negative, the curves 6.2 and 6.4 were generated. From these graphs it is possible to conclude that the Suvrel method has higher AUCs in most of cutoffs. It is appropriate to note that conclusions were obtained using a method that does not make any assumption about the distribution associated with the reads, using a simple normalization (divide the counts of a gene by its standard deviation or setting them having zero mean and variance 1) with a fast execution.

Keywords: RNA-Seq. Differential expression. Suvrel method. Analysis.

Lista de Figuras

1.1	Figura que relaciona o custo de sequenciamento do genoma humano no decorrer dos anos.	14
4.1	Fluxograma dos passos envolvidos em uma análise para detecção de expressão diferencial de genes.	19
4.2	Passos envolvendo RNA-Seq através de cDNA	21
4.3	Esquema representativo da plataforma Helicos de sequenciamento. . . .	23
5.1	Curva ROC de um conjunto de dados de exemplo gerado por um classificador randômico.	37
6.1	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do \log_2 do <i>fold-change</i> de 0,5 e normalização pelo desvio padrão.	47
6.2	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do \log_2 do <i>fold-change</i> de 0,5 a 2,0 e normalização pelo desvio padrão. . . .	48
6.3	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do \log_2 do <i>fold-change</i> de 0,5 e normalização média zero e variância 1. . . .	48
6.4	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do \log_2 do <i>fold-change</i> de 0,5 a 2,0 e normalização média zero e variância 1. . . .	49
6.5	Análise ROC a partir do conjunto de dados ERCC usando o log predefinido das proporções das mistura ERCC 1 e ERCC 2 para definição dos verdadeiros positivos e negativos, e normalização pelo desvio padrão. . . .	50
6.6	Análise ROC a partir do conjunto de dados ERCC usando o log predefinido das proporções das mistura ERCC 1 e ERCC 2 para definição dos verdadeiros positivos e negativos, e normalização média zero e variância 1. . . .	51

6.7	Análise ROC-like a partir do conjunto de dados ERCC usando o log predefinido das proporções das mistura ERCC 1 e ERCC 2, normalização pela média zero e variância 1 e distância relacionando um par de genes.	52
6.8	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do <i>fold-change</i> de 0,5, normalização pela média zero e variância 1 e distância relacionando um pares de genes.	53
6.9	Análise ROC-like a partir do conjunto de dados TaqMan, usando limiar do <i>fold-change</i> de 0,5 a 2,0 e normalização pela média zero e variância 1.	54
6.10	Análise ROC a partir do conjunto de dados TaqMan, usando limiar do \log_2 do <i>fold-change</i> de 0,5, normalização pela média zero e variância 1, distância relacionando pares de genes e as relevâncias pelas equações 5.30, 5.31, 5.32.	56
6.11	Análise ROC a partir do conjunto de dados ERCC usando o log predefinido das proporções das mistura ERCC 1 e ERCC 2, normalização pela média zero e variância 1, distância relacionando pares de genes e as relevâncias calculadas pelas equações 5.30, 5.31, 5.32.	57
7.1	Gráfico em barras mostrando os valores abaixo da curva ROC, a partir do conjunto de dados ERCC usando o log predefinido das proporções das mistura ERCC 1 e ERCC 2, normalização pela média zero e variância 1 e distância relacionando um par de genes.	59

Lista de Tabelas

4.1	Tabela informativa que lista os métodos de normalização dos pacotes.	28
4.2	Tabela comparativa dos métodos de modelagem estatística da expressão de genes dos pacotes.	32
4.3	Tabela comparativa dos métodos de normalização dos pacotes.	33
5.1	Tabela de <i>scores</i> de um classificador probabilístico hipotético.	36
5.2	Tensor métrico hipotético para exemplo.	42
5.3	Tensor métrico hipotético de exemplo usado para mostrar como são obtidas as relevâncias usando a diagonal principal.	43
5.4	Tensor métrico hipotético para exemplo usado para mostrar como é obtida a relevância quando ela depende do gene e das associações que ele pode fazer.	43
5.5	Tensor métrico hipotético para exemplo usado para mostrar como é obtida a relevância de um gene dependendo da média das associações em pares que ele que pode realizar.	44

Sumário

1	Introdução	13
2	Motivação	16
3	Objetivos	17
4	Experimentos de RNA-Seq para verificação de expressão diferencial	18
4.1	Visão geral dos passos envolvidos na análise de expressão diferencial . .	18
4.2	Sequenciamento de RNA	20
4.2.1	RNA-Seq usando síntese de cDNA	20
4.2.2	Sequenciamento direto de RNA	20
4.3	Mapeamento ou alinhamento	22
4.4	Sumarização	24
4.5	Normalização	25
4.6	Distribuições estatísticas e Modelagem estatística da expressão de genes	29
4.6.1	Distribuições de Poisson	29
4.6.2	Distribuições binomial negativa	29
4.6.3	Modelagem estatística da expressão de genes	30
4.7	Teste para expressão diferencial	32
5	Metodologia	34
5.1	Curvas <i>ROC</i>	34
5.2	R estatística	38

5.3	Pacotes utilizados neste estudo para inferência de expressão diferencial	38
5.4	Uso do método Suvrel na inferência de expressão diferencial em dados originados de RNA-Seq	38
5.5	Conjunto de dados usado para análise	44
6	Resultados e Discussão	46
7	Conclusões e Perspectivas Futuras	58
	Referências	58
A	Multiplicadores de Lagrange	63
B	Obtenção da equação 5.12 utilizando a distância euclidiana	65
C	Obtenção da expressão da relevância usando a equação 5.21	67

Capítulo 1

Introdução

Este milênio está sendo marcado pela recente revolução científica que teve um grande impacto na sociedade, assim como as revoluções industrial e tecnológica tiveram nos séculos passados. Essa nova revolução teve fato marcante em julho de 1995, após o sequenciamento completo do genoma da bactéria *Haemophilus influenzae* que continha 1,8 milhões de pares de bases [1]. Alguns anos depois, o Projeto Genoma Humano, que é um consórcio internacional de 16 instituições, anunciou em Junho de 2001 um esboço do sequenciamento do DNA humano [2]. Em uma iniciativa privada paralela ao Projeto Genoma Humano, a companhia Celera, sob o direcionamento de J. Craig Venter também anunciou feito igual no mesmo ano [3]. Desde então, a quantidade de DNA sequenciado e depositado em bibliotecas públicas cresce exponencialmente [4, 5, 6]. Entre as razões para este crescimento [1], podem-se elencar inúmeros fatores mas a diminuição considerável do custo de sequenciamento, que pode ser vista na Figura 1.1, pode resumir a argumentação para explicar essa intensificação. Essa diminuição do custo de sequenciamento teve um aumento pronunciado a partir dos anos 2006 e 2007 devido ao advento do sequenciamento de alto desempenho (*Next-Generation Sequencing, NGS*) que por sua vez trouxe um aumento considerável no volume de dados gerados em cada experimento.

A partir dessa mudança significativa na geração de dados, a hipótese reducionista pré-genômica da Biologia Molecular, que era aplicada em sistemas pequenos, foi modificada para a abordagem pós-genômica, que é aplicada em grandes sistemas e dirigida pela quantidade massiva de dados, atualmente na escala de armazenamento e processamento em terabytes [4]. Essa mudança trouxe a necessidade de métodos quantitativos estatísticos e computacionais de alta performance para análise de dados de sistemas biológicos. Sendo assim, dada a grande quantidade de informação gerada, a situação atual não é somente produzir, mas compreender os dados por meio



Figura 1.1: Figura que relaciona o custo de sequenciamento do genoma humano no decorrer dos anos. Adaptado de: <http://www.technologyreview.com/graphiti/427720/bases-to-bytes/> de ferramentas de bioinformática [7].

Entre as modalidades de sequenciamento que foram impactadas pela NGS, pode-se atentar ao RNA-Seq que, com ajuda dos algoritmos de montagem, tornou possível a reconstrução de transcritomas completos com resolução melhor que de microarranjos [8]. Com as informações deste tipo de sequenciamento é possível também a análise de transcritos inteiros, ajudando a desvendar a dinâmica dos transcritomas [7, 9, 10]. Além da reconstrução de transcritoma, o avanço da tecnologia de RNA-seq tem propiciado outros desafios e oportunidades para pesquisadores, como a quantificação e análise de transcritos expressos diferencialmente [9].

As informações geradas através do uso das tecnologias ligadas à expressão de genes são usadas com o objetivo de fornecer meios para aumentar o entendimento da atividade transcricional em diferentes tecidos ou populações de células, por meio da identificação das mudanças de expressão de transcritos associados à fenótipos de interesse ou condições de tratamento (amostras ligadas a alguma doença *versus* amostras saudáveis). Essas informações podem vir de um estudo onde um sistema biológico é perturbado pela inativação de um gene, onde pesquisadores podem estar interessados em saber se um gene particular facilita ou bloqueia a ação de uma determinada droga, comparando amostras onde este gene foi inativado em amostras normais *versus* amostras sob efeito de tais drogas. Essas informações podem ainda vir através da aplicação de um fator de estresse, podendo fornecer *insights* sobre processos celulares normais ou ligados à alguma doença, ou ainda podem vir de estudos observacionais comparando, por exemplo, tecido normal *versus* doente ou células de diferentes populações [11, 12].

Entre essas tecnologias, o RNA-Seq também pode contribuir neste âmbito por meio da análise de expressão diferencial de um dado gene a partir da tabela de contagens (que sumariza cada biblioteca em uma dada condição experimental), mostrando o transcrito que teve sua abundância significativamente alterada entre as condições experimentais. Dependendo do objetivo do estudo, a geração da lista de genes não é a conclusão final, mas um passo na análise e, sendo assim, o estudo tem o objetivo de um *insight* biológico que pode ser feito olhando mudanças de expressão de um conjunto de genes. O objetivo, contudo, pode ainda ser mais complexo e os dados gerados por RNA-Seq, como a lista de genes expressos diferencialmente, pode ser integrada com outras fontes de informações, como por exemplo, estudo de RNA de interferência, modificação de histona e metilação de DNA, para estabelecer um entendimento mais claro sobre mecanismos regulatórios [8].

No que diz respeito a análise de expressão diferencial, ela se baseia na identificação dos transcritos que apresentaram mudanças estatisticamente significantes na abundância de entre diferentes condições além daquela que é esperada randomicamente [8]. Dois tipos de variações podem ser distinguidos em qualquer experimento RNA-Seq. O primeiro tipo é aquele que está ligado à variação relativa devido a causas biológicas e o segundo tipo está ligado à incerteza com o qual a abundância do transcrito é estimada através da tecnologia de sequenciamento [13]. Portanto, um transcrito é declarado diferencialmente expresso se as quantidades diferem significativamente (atentando para os dois tipos de variações) em grupos distintos de amostras. Tal avaliação utilizando RNA-Seq, inicia-se com o sequenciamento das amostras, gerando bilhões de pequenas seqüências que passam pelo *pipeline*, que se inicia com o mapeamento, sumarização, contagem e normalização para posterior análise estatística que apontará o transcritos ou conjunto de transcritos que tiveram uma mudança estatisticamente relevante em suas abundâncias. Embora existam pesquisas em todos esses passos, ainda existe a necessidade de aperfeiçoamentos e melhorias, onde tais refinamentos podem ir desde maiores estudos sobre qual a métrica para a sumarização mais adequada até qual o efeito da inflexibilidade das suposições (é comum os pacotes destinados a expressão diferencial assumirem que as contagens seguem algum tipo de distribuição) feitas pelos métodos de inferência de expressão diferencial [8].

Capítulo 2

Motivação

O uso de sequenciamento de alto rendimento produz um volume denso de informações, impondo a necessidade de metodologias adequadas que ajudem pesquisadores e correlatos a interpretar de maneira apropriada as informações [8]. O sequenciamento de RNA de nova geração também é integrante deste cenário e os pacotes e as metodologias usadas ainda carecem de melhorias e aprimoramentos [8]. Dada a necessidade de novas ferramentas, propõe-se a utilização de um método variacional para análise de expressão diferencial de genes a partir de uma tabela de contagens gerada com dados de RNA-Seq. Tal proposta é motivada por estudos anteriores que mostraram que a metodologia chamada de Aprendizado de Relevância Variacional Supervisionado (*Supervised Variational Relevance Learning, Suvrel*) melhorou a performance de ferramentas computacionais de predição de dados originados de microarranjos e concentrações de metabólitos medidos por ressonância magnética nuclear [14].

Capítulo 3

Objetivos

Dada a motivação exposta no capítulo 2, é proposto o uso do tensor métrico calculado através do método Suvrel para ranqueamento de genes envolvidos em expressão diferencial a partir da tabela de contagens gerada por dados de RNA-Seq.

A partir desta proposta, o primeiro objetivo é a comparação da implementação do método com outras ferramentas públicas. Tal estudo será feito usando curvas ROC (*Receiver Operating Characteristic*) a partir de uma tabela de contagens que sumariza 10 amostras de RNA disponibilizada publicamente por [15], e dados do projeto MAQC-II (*MicroArray Quality Control*) e ERCC (*External RNA Control Consortium*). As amostras de RNA são constituídas de 5 amostras, que caracterizam o grupo A, contendo RNAs de referência humano universal juntamente com 2% de volume da mistura ERCC 1, e 5 amostras, que caracterizam o grupo B, contendo RNAs de referência do cérebro humano juntamente com 2% de volume da mistura ERCC 2. As informações disponibilizadas pelo projeto MAQC-II, em forma de arquivo texto, relacionam a abundância de cerca de 1000 genes medidas por TaqMan qPCR em amostras de RNA humano do tecido cerebral e RNA de diferentes tecidos, e as informações disponibilizadas pelo ERCC, por meio de arquivo texto, são ligadas as proporções predefinidas das misturas ERCC 1 e ERCC 2, formando o conjunto utilizado para definir os RNA que são usados para definir os verdadeiros negativos e positivos na produção das curvas.

Após a conclusão da primeira etapa, o segundo objetivo é disponibilizar a implementação do método, chamada de DESuvrel, para a comunidade acadêmica.

Capítulo 4

Experimentos de RNA-Seq para verificação de expressão diferencial

4.1 Visão geral dos passos envolvidos na análise de expressão diferencial

Na Figura 4.1 pode ser visto, de uma forma geral, que um experimento de RNA-Seq de análise de expressão diferencial se inicia com o sequenciamento das amostras, gerando de milhões a bilhões de pequenas sequências de nucleotídeos (*reads*) [16] que passarão por verificação de qualidade para retirada de artefatos e, se necessário, alguns podem ser descartados caso a qualidade seja baixa [8]. Após o mapeamento dos *reads*, eles são sumarizados e agregados a partir de uma unidade de significado biológico (por exemplo, genes, transcritos ou exons). Posteriormente a esta etapa, gera-se uma tabela de contagens que deverá ser normalizada utilizando uma abordagem atrelada ao método estatístico de teste que será usado para a inferência da expressão diferencial, como acontece também com a implementação do método Suvrel (chamada de DESuvrel) que terá duas normalizações associadas que serão analisadas. Os passos anteriores são seguidos pela aplicação da abordagem de inferência, como por exemplo as presentes nos pacotes edgeR, DESeq e o DESuvrel, que por sua vez gerará uma lista de genes (ou qualquer outra característica genômica de interesse) acompanhados pelos respectivos *fold-changes* e valores-p, na maioria dos casos. Essa lista é então utilizada em abordagens ligadas à biologia de sistemas para culminar em um *insight* biológico [8]. Dada as nuances e características próprias de cada passo deste tipo de estudo, os tópicos a seguir tratarão com mais detalhes cada um deles.

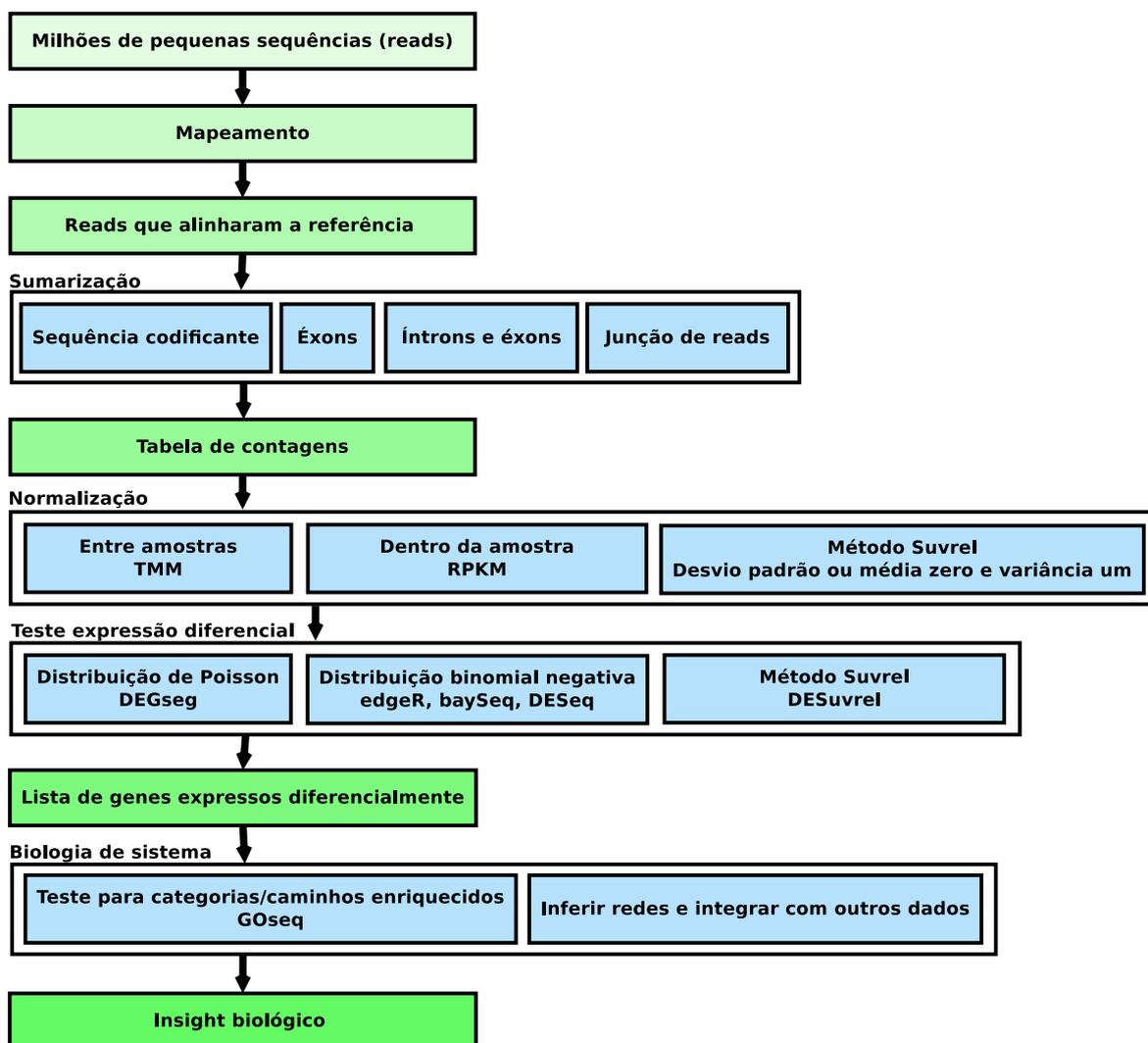


Figura 4.1: Fluxograma dos passos envolvidos em uma análise para detecção de expressão diferencial de genes. Milhões de *reads* são produzidos após amostras de RNA terem sido sequenciadas (caixa Milhões de pequenas sequências (*reads*)). O mapeamento pode ser produzido com o auxílio de um genoma ou transcrito de referência usando por exemplo o programa bowtie (caixa Mapeamento). Após os *reads* terem sido alinhados, a próxima etapa é a sumarização, que basicamente conta uma característica ligada ao objetivo do estudo, que no caso mais geral são os *reads* que alinharam aos respectivos genes de origem (caixa Sumarização). A partir da conclusão desta etapa, tem-se gerada a tabela de contagens que deverá ser normalizada (caixa Tabela de contagens e Normalização). A inferência de expressão diferencial é executada no teste de expressão diferencial, podendo ser realizada supondo que os *reads* seguem uma distribuição de Poisson, binomial negativa ou utilizando o método Suvrel que não necessita de nenhuma suposição dessa natureza (Caixa Teste de expressão diferencial). Na conclusão desta etapa tem-se gerada a lista de genes diferencialmente expressos (Caixa Lista de genes expressos diferencialmente). Os passos envolvidos na etapa chamada de biologia de sistemas são os estudos dos resultados obtidos com os passos anteriores, juntamente com outros dados auxiliares no sentido de obter um *insight* associado ao objetivo que motivou o estudo. Adaptado de [8].

4.2 Sequenciamento de RNA

A tecnologia de microarranjo destinada exclusivamente a análise de expressão diferencial foi uma ferramenta amplamente usada na Biologia, mas recentemente está sofrendo declínio na sua utilização devido à chegada de tecnologias de alto rendimento de sequenciamento de DNA que tornam possível o sequenciamento de RNA em grande escala através de DNA complementar. Atualmente a necessidade da utilização de cDNA não é mais obrigatória mas as vantagens sobre os microarranjos permanecem, como, por exemplo, a baixa taxa de erro e a não necessidade de conhecimento prévio do organismo a ser sequenciado [9, 17].

Empresas como Illumina, Roche 454, Helicos BioSciences e Life Technologies comercializam equipamentos para RNA-Seq e também investem em novas tecnologias [9]. E assim dada tal dinâmica, serão descritos somente sobre um exemplo de sequenciamento utilizando cDNA e outro por síntese.

4.2.1 RNA-Seq usando síntese de cDNA

Na Figura 4.2 é possível visualizar que os passos típicos envolvidos no protocolo deste tipo de sequenciamento se iniciam com a fragmentação do RNA total ou mRNAs que, posteriormente, são convertidos em uma biblioteca de cDNAs contendo adaptadores de sequência. A biblioteca de cDNA é então sequenciada para produzir de milhões a bilhões de pequenas sequências (*reads*) a partir de um ou de ambos os finais dos fragmentos de cDNA [7].

4.2.2 Sequenciamento direto de RNA

Para ilustrar os passos fundamentais envolvidos no sequenciamento de RNA por síntese ilustrados na Figura 4.3, a abordagem desenvolvida pela Helicos será usada como exemplo. Neste aparelho, a captura pela cauda 3' dos nucleotídeos poli-A que serão sequenciados é feita por oligonucleotídios poli-dT que são ligados covalentemente pela cauda 5' à uma superfícies de vidro ultra-limpo. O sequenciamento se inicia em um passo conhecido como “preencher”, integrante do par de etapa “preencher e travar”, que tem objetivo de expor os pares hibridizados de RNA e poli-dT à polimerases e um excesso de timinas naturais, para que o sequenciamento não inicie na cauda poli-A. No passo “travar”, nucleotídeos adenina, citosina e guanina contendo grupos fluorescentes e quimicamente cliváveis são hibridizados às respectivas bases complementares, fechando esta fase. O processo é seguido por uma etapa de lavagem, onde os nucleotídeos jun-

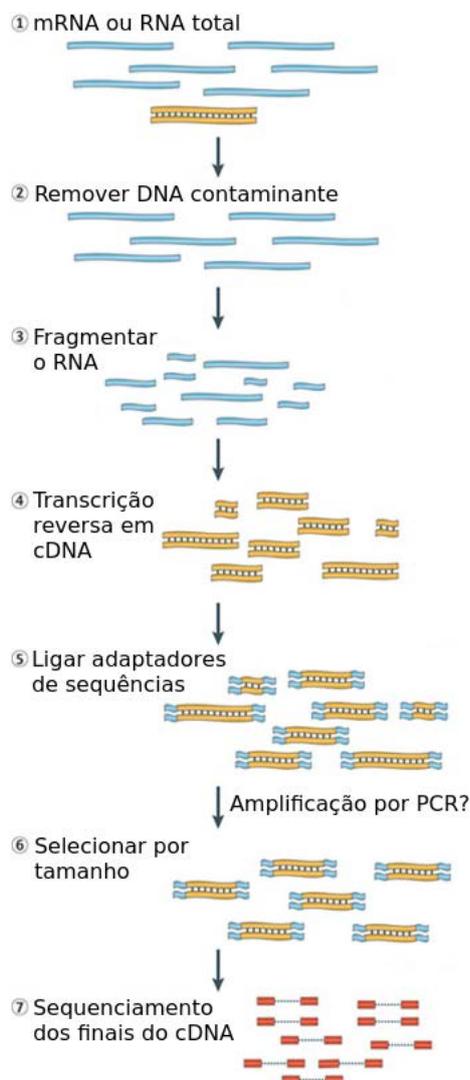


Figura 4.2: A utilização do RNA-Seq através de cDNA se inicia com a fragmentação do RNA total ou mRNAs que foram previamente extraídos (estágio 1). Em seguida o DNA contaminante é removido usando DNase (estágio 2) e no final do terceiro estágio o RNA está fragmentado em pequenas frações (estágio 3). Os fragmentos de RNA são então transcritos reversamente em cDNA (estágio 4, amarelo) e em seguida adaptadores de sequência (azul) são ligados (estágio 5) e posteriormente os fragmentos são selecionados por tamanho (estágio 6). Finalmente, os finais dos fragmentos dos cDNAs são então sequenciados. É importante notar que no quinto passo é possível amplificar as amostras, e uma alternativa deste protocolo pode gerar o sequenciamento de somente um final de cada fragmento. Adaptado de [7].

tamente com o excesso de timina que não foram incorporados são descartados. Este ciclo de sequenciamento termina quando imagens são geradas e uma mistura que cliva o corante fluorescente do último nucleotídeo hibridizado é adicionada, tornando as fitas adequadas para outra ciclo de incorporação. A reação de sequenciamento continua com a adição dos próximos nucleotídeos com grupos corantes e cliváveis seguida pela lavagem, geração de imagens e clivagem. A repetição deste ciclo por muitas vezes, fornece um grande conjunto de imagens onde as incorporações das bases são detectadas e então usadas para gerar a informação de sequenciamento para cada *read* individual. Este equipamento contém cerca de 50 canais independentes, e sendo assim cada utilização produzirá entre 800.000 a 12.000.000 *reads* alinhados, com tamanho de 25 a 55 e média de 33 nucleotídeos, dependendo do tempo de execução e da qualidade da geração das imagens [18, 9] (devido as tecnologias de sequenciamento continuarem em desenvolvimento estes números provavelmente continuarão a crescer).

4.3 Mapeamento ou alinhamento

Na etapa de um experimento de expressão diferencial conhecida como mapeamento ou alinhamento é onde os milhares de *reads* são convertidos em medidas de quantificação de expressão [8, 19], sendo, portanto, uma parte crucial no *pipeline* das análises ligadas a RNA-Seq [16].

Para a realização do mapeamento existem inúmeras ferramentas [16], onde elas podem ser divididas em duas categorias principais: as ferramentas baseadas em tabelas *hash* e as baseadas na transformação *Burrows–Wheeler* [20]. Uma tabela *hash* é uma estrutura de dados, usada para indexação de conjuntos de dados complexos para facilitar a busca rápida por *strings*. *Burrows–Wheeler* é um rearranjo reversível de *strings* que codifica o genoma em uma representação mais compacta, produzindo redundâncias de subsequências repetidas [19].

Embora à primeira vista a tarefa de encontrar uma localização única onde um *read* possua sequência idêntica a referência pareça ser fácil, isto não acontece na prática [8]. A primeira consideração à respeito é obtida ao observar que a referência nunca representa, de maneira perfeita, a fonte biológica do RNA sendo sequenciado, ou seja, o RNA sequenciado e estudado não foi o utilizado na produção da referência utilizada no alinhamento. Somado a isto, tem-se atributos específicos da amostra, como a variação na sequência de DNA que afeta somente uma base (*Single Nucleotide Polymorphism*, *SNP*) e *indels* (inserções ou deleções), e também a consideração de que as sequências se originam de transcritos que vieram de *splicing* mais do que de

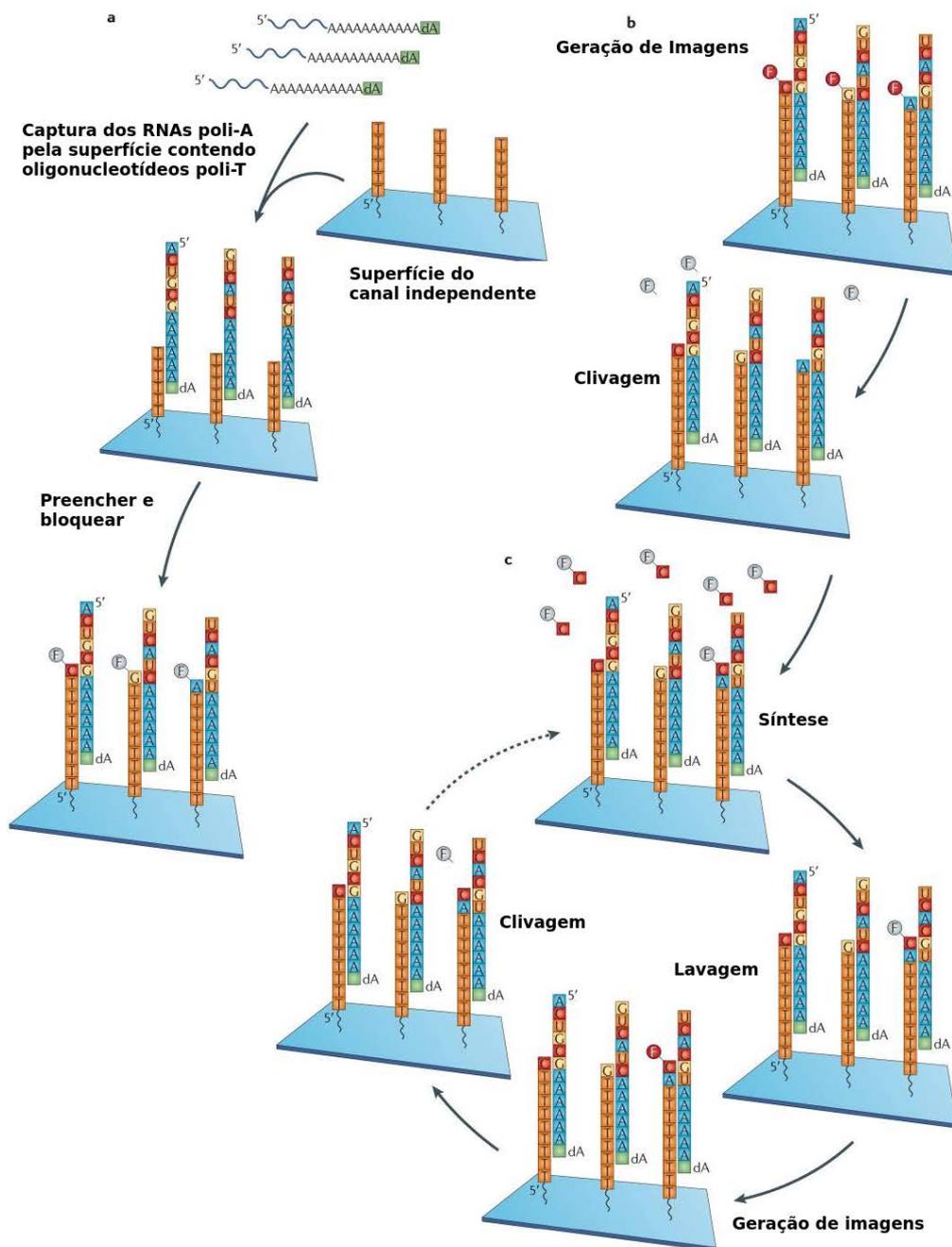


Figura 4.3: Esquema representativo da plataforma Helicos de sequenciamento. RNA poli-A com a 3' é capturado pela superfície contendo poli-dT. O passo posterior envolve uma etapa de "fill-and-lock" que introduz nucleotídeos A, C e G marcados quimicamente. Estes últimos passos garantem que o sequenciamento ocorra de maneira correta de forma que a cauda poli-A não é erroneamente sequenciada. Na parte b, imagens são geradas e o fluoróforo das bases que foram ligadas complementarmente é clivado. Na parte c, ciclos de introdução de bases ligadas à fluoróforos, geração de imagens e clivagens, são executados. Ao final do sequenciamento, são produzidas entre 800.000 a 12.000.000 de pequenas sequências, com tamanho de 25 a 55 e média de 33 nucleotídeos dependendo do tempo de execução e da qualidade das gerações das imagens (as tecnologias de sequenciamento continuam em desenvolvimento e provavelmente estes números são maiores). Fonte: [18].

um genoma, produzindo uma situação complexa de ser analisada. Além disso, as sequências podem se alinhar perfeitamente a múltiplas localizações e podem conter erros de sequenciamento que tem de ser levados em conta [8, 19]. Quando considera-se *reads* obtidos da transcrição reversa de DNA, onde um final pode ser sequenciado (tópico 4.2.1) as ferramentas de alinhamento diferem em como eles manipulam *reads* que mapearam igualmente em várias regiões. Entre as opções que um algoritmo pode utilizar é possível descartar tais *reads*, alocá-los randomicamente ou utilizar métodos estatísticos que incorporam *scores* de alinhamento. Este problema é mais evidente quando somente um final dos fragmentos é sequenciado e a utilização dos dois finais deve amenizar esta situação [8].

Quase todas as ferramentas de alinhamento usam a estratégia de primeiro executar um *matching* eurístico que rapidamente gera uma lista de possíveis localizações, seguida por uma avaliação completa de todos os candidatos de alinhamento por um complexo algoritmo de alinhamento local exato [8, 19]. Embora exista pesquisa na análise da melhor forma de alinhar as sequências, todas as soluções envolvem necessariamente algum consenso entre os recursos computacionais utilizados pelos algoritmos e a incerteza permitida (vinda por exemplo ao ignorar informação de SNPs, negligenciar os *scores* de qualidades das bases, limitação do número de *mismatch* permitidos) no *matching* à referência [16].

4.4 Sumarização

Dada a dificuldade exposta na seção 4.3 anterior, nesta fase do experimento, a partir da obtenção das localizações das sequências o tanto quanto foi possível, elas são sumarizadas sobre alguma unidade de significado biológico, como por exemplo, genes, éxons ou transcritos, por meio das ferramentas HTSeq, Picard, BEDTools entre outras [16, 8].

Entre as variações possíveis desta fase, a mais simples é contar o número de sequências que sobrepõem éxons de genes. Contudo, este pode não ser o objetivo proposto, ou ainda exista uma proporção significativa das sequências que mapeiam em regiões genômicas fora dos exons anotados, até mesmo em organismos bem anotados [19, 21]. Para solucionar este problema, pode-se incluir, na sumarização, as sequências que alinham em todo o comprimento do gene e, naturalmente, incorporar íntrons na contagem, e ainda colateralmente incluir éxons pobremente ou não anotados assim como fronteiras de éxons variáveis. A partir do panorama descrito, é possível constatar que nesta etapa existem muitas opções, mas é obrigatório ter atenção na escolha da

sumarização pois ela tem o potencial de mudar substancialmente as contagens de cada gene e levar a conclusões errôneas [8, 16].

Consequentemente, após a conclusão desta etapa é gerada uma matriz de contagem N de n por m , onde N_{ij} é o número de sequências atribuídas ao gene i no experimento de sequenciamento j , n é o número de genes estudados e m é o número de bibliotecas. Neste ponto, é importante salientar que o número de sequências que se sobrepõem a um dado gene i não é uma medida direta da expressão do gene. A medida de contagem N_{ij} é proporcional a $l_i \mu_{ij}$, onde μ_{ij} e l_i são, respectivamente, a expressão esperada (número de fragmentos associados ao gene i) e o comprimento do gene, caracterizando um viés claro de comprimento. Um produto deste viés é reduzir a habilidade de detecção de expressão diferencial entre genes com contagens pequenas devido a falta de cobertura e a repercussão deste efeito deve ser analisado com cautela no passo seguinte de normalização [15].

4.5 Normalização

Nesta fase, as diferenças entre números de sequências entre diferentes corridas de sequenciamento, presença de viés técnicos introduzidos por protocolos de preparação de bibliotecas assim como de diferentes plataformas de sequenciamento devem ser contabilizadas. Portanto, o objetivo da normalização é tentar resolver ou amenizar tais efeitos indesejados para facilitar comparações acuradas entre as amostras [15].

Para que seja possível discorrer mais detalhadamente sobre este tema, considere dois experimentos de RNA-Seq sem genes expressos diferencialmente. Contudo, se imaginarmos que o primeiro experimento gere duas vezes mais sequências que o segundo, é natural concluir que as contagens do primeiro experimento são duplicadas e, assim, sermos inclinados a dizer que ocorre expressão diferencial. Este cenário, que pode levar a conclusões errôneas, acontece devido a diferentes profundidades de sequenciamento d_j (do inglês *sequencing depth*, sendo definido como o número total de *reads* sequenciados ou mapeados) e é o primeiro viés que será discutido nesta seção. Assim, os primeiros trabalhos direcionados a corrigir este problema utilizaram uma técnica que ficou conhecida como escalonamento global, onde as contagens em cada experimento j eram escalonadas por d_j [8, 15, 19].

Contudo, existem algumas abordagens que se atentam mais em amenizar o viés existente devido a predominância de um conjunto restrito de genes altamente expressos, responsáveis por consumirem boa parte dos *reads* disponíveis, deixando alguns genes subestimados. Na intenção de corrigir este problema colateral, Bullard

et al. propuseram uma normalização por quantil, implementada no pacote baySeq [22], na forma de um escalonamento global que ajusta as distribuições das contagens considerando seus terceiros quartis. Seguindo este mesmo objetivo, Robinson e Oshlack et al. propuseram o método de normalização conhecido como “*Trimmed Mean of M-values (TMM)*”, implementado no edgeR [23]. Para reduzir este viés, o TMM utiliza a estratégia de calcular um fator de normalização após remover 30% dos genes que tem os mais extremos logs de *fold-changes*, conhecidos como valores-M, para então usar este fator para corrigir as diferenças nos tamanhos das bibliotecas [19, 15].

Para elucidar detalhadamente o método de normalização utilizado pelo pacote DESeq [24], faz-se necessário inicialmente estabelecer N_{ij} como a contagem atribuída ao gene i no experimento j . A partir desta definição, se considerarmos duas replicatas atribuídas com $j=1$ e $j=2$, é esperado que um histograma das proporções

$$\frac{N_{i1}}{N_{i2}} \quad (4.1)$$

considerando todos os genes i mostre um pico bem definido, e que sua mediana represente uma boa estimativa da proporção da diferença na profundidade de sequenciamento. Como na maioria dos estudos existem mais do que duas replicatas, uma amostra de referência é produzida calculando a média geométrica f_i de um gene i considerando todas as amostras até ter este cálculo para todos os genes. Assim, o fator de escalonamento da amostra j é obtido a partir da mediana

$$s_j = \frac{N_{ij}}{f_i} \quad (4.2)$$

considerando todos os genes i . A partir da obtenção do valor de s_j , divide-se cada contagem por esse valor para se obter as contagens normalizadas. O pacote Cuffdiff estende estes passos realizando, primeiramente, um escalonamento entre as mesmas condições experimentais e, depois, entre as condições diferentes [19, 15, 24].

Um dos pacotes que também será analisado neste trabalho é o PoissonSeq [25], que utiliza uma estimativa vinda de um ajuste de qualidade (*goodness-of-fit*) para definir um conjunto de genes que é pelo menos diferenciável entre duas condições para então calcular os fatores de normalização [15].

O pacote limma [26] desenvolvido para análise de expressão diferencial utilizando dados vindos de microarranjo, utiliza normalização por quantil. Recentemente este pacote ganhou uma nova função de normalização chamada *voom* [27], desenvolvida especificamente para dados de RNA-Seq. De uma forma bastante geral, este método realiza uma regressão *LOWESS* para estimar a relação entre a média e a variância e

transforma as contagens na escala log para executar uma transformação linear [15].

Esta seção, assim como as anteriores, continua em evolução, mas entre tantas variações, os métodos TMM e o utilizado pelo DESeq tem sido mais efetivos como apontado por diferentes estudos [19]. Contudo, se a suposição de que as amostras não contém quantidade equiparada de RNA, tais métodos não geraram resultados satisfatórios e outras técnicas relacionadas à medidas de RNA *spike-in* podem ser usadas [19].

Condizente com a fervorosa discussão quanto aos métodos de normalização, e no sentido de contribuir para elucidação de uma solução para estes entraves, este propõe a utilização de duas formas de normalização. Na primeira forma (referenciada no texto como “**média 0 e variância 1**”), as contagens k_i do gene i serão subtraídas da média e divididas pelo desvio padrão, como pode ser visto na equação 4.3.

$$k_{Ni} = \frac{k_i - \mu}{\sigma_i} \quad (4.3)$$

onde k_{Ni} é a contagem normalizada, k_i é a contagem original, μ é a média das contagens e σ_i é o desvio padrão. Pode existir uma situação onde as contagens são as mesmas entre as condições, e o cálculo do σ_i será 0 e não poderá ser usado na normalização. Para resolver este obstáculo, a implementação do método está preparada para utilizar o menor desvio padrão, ou seja, o método procura o menor desvio padrão existente entre os genes e atribui este valor ao desvio padrão daquele que tem valor 0. No segundo tipo de normalização (referenciada como “**normalização pelo desvio padrão**”), as contagens k_i do gene i serão divididas pelo desvio padrão σ_{gi} da seguinte maneira:

$$k_{Ni} = \frac{k_i}{\sigma_{gi}} \quad (4.4)$$

$$\sigma_{gi} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \quad (4.5)$$

onde σ_A^2 é a variância das contagens do gene i dentro de uma dada condição experimental A , n_A é o número de contagens dentro da mesma condição A , e o segundo termo da expressão é análogo, respeitando a adequação a uma outra condição experimental B . No cálculo foram consideradas duas condições experimentais A e B , mas no caso de existir uma terceira condição C é necessário somente adicionar um termo ao cálculo da equação 4.5 como segue:

$$\sigma_{gi} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} + \frac{\sigma_C^2}{n_C}} \quad (4.6)$$

Tabela 4.1: Tabela informativa que lista os métodos de normalização executados pelos pacotes que serão estudados juntamente com as informações do DESuvrel.

Pacote	Método
baySeq	Escalonamento global que ajusta as distribuições das contagens de um experimento a partir do terceiro quantil.
edgeR	Escalonamento global utilizando um fator que foi calculado após 30% dos genes com os mais extremos logs de <i>fold-changes</i> terem sido retirados.
DESeq	Utiliza escalonamento global a partir de um fator s_j que é calculado como a mediana de $\frac{N_{ij}}{f_i}$, onde N_{ij} é a contagem atribuída ao gene i no experimento j e, f_i é a média geométrica de um gene i considerando todas as amostras
CuffDiff	Os passos executados pelo DESeq são estendidos, onde é realizado primeiramente um escalonamento entre as mesmas condições experimentais e depois entre as condições diferentes.
PoissonSeq	Utiliza uma estimativa vinda de um ajuste de qualidade para definir um conjunto de genes que é pelo menos diferenciável entre duas condições para então calcular os fatores de normalização das bibliotecas.
limmaQN	Normalização por quantil.
limmaVoom	Realiza uma regressão <i>LOWESS</i> para estimar a relação entre a média e a variância e transformar as contagens na escala log para obter uma transformação linear.
DESuvrel	Normalização aplicada por gene utilizando o desvio padrão ou média zero e variância 1.

onde σ_C^2 é a variância das contagens do gene i dentro de uma dada condição C , n_C é o número de contagens dentro da mesma condição C . No caso da existência de outra condição a ideia por trás do cálculo continuará a mesma. Assim como no primeiro tipo de normalização, em uma determinada situação um gene pode possuir um dado número de contagens iguais nas replicatas da condição experimental A e outro número igual nas contagens dentro da condição B. Neste caso, o cálculo do desvio padrão por meio da equação 4.5 resultará no valor 0, e ao dividir as contagens do gene por este valor será gerada uma inconsistência matemática. Assim, quando ocorrer esta situação o valor de σ_{gi} será trocado pelo menor desvio padrão como já citado anteriormente.

As informações resumidas sobre qual o método de normalização executado pelos pacotes, incluindo o DESuvrel, que serão estudados, são encontradas na tabela 4.1.

4.6 Distribuições estatísticas e Modelagem estatística da expressão de genes

4.6.1 Distribuições de Poisson

A distribuição de Poisson é amplamente utilizada, podendo servir como modelo para um amplo número de diferentes tipos de assuntos que envolvam fenômenos que estamos esperando por uma ocorrência, como por exemplo, esperar por um ônibus ou chegadas de clientes em um banco. Para que seja possível o uso desta distribuição, é necessário que a suposição na qual ela foi construída seja satisfeita, ou seja, a probabilidade de uma chegada, utilizando o exemplo do banco, seja proporcional ao tamanho do tempo de espera [28]. Uma variável randômica x , tomando valores inteiros não negativos, tem uma distribuição de Poisson se

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad (4.7)$$

onde $x = 0, 1, \dots$, e o parâmetro λ , é algumas vez chamado de parâmetro de intensidade [28]. A média de X é igual ao valor esperado, onde este último é dado por [28]

$$E(X) = \lambda \quad (4.8)$$

A variância é dada por [28]

$$Var(X) = \lambda \quad (4.9)$$

4.6.2 Distribuições binomial negativa

A distribuição binomial conta o número de sucessos numa sequência de n tentativas, se as tentativas forem independentes, onde cada tentativa resulta apenas nas possibilidades de sucesso ou fracasso (tentativas de Bernoulli). A partir da distribuição binomial é possível obter a distribuição binomial negativa, onde a partir de uma sequência de tentativas independentes de Bernoulli, a variável X , que denota a tentativa na qual o r -ésimo sucesso ocorre, é dita seguir uma distribuição binomial negativa (r, p) se:

$$P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad (4.10)$$

onde $x = r, r + 1$ e r é um inteiro fixo [28].

Existem casos onde existe a necessidade de definir a distribuição binomial

negativa em termos da variável randômica $Y = \text{número de falhas antes do } r\text{-ésimo sucesso}$. Esta formulação é estatisticamente equivalente aquela dada em termos de $X = \text{número de tentativas no qual o } r\text{-ésimo sucesso ocorre}$, desde que $Y = X - r$. Assim, usando a relação entre Y e X ($Y = X - r$), a distribuição binomial também pode ser escrita como [28]:

$$P(Y = y) = (-1)^y \binom{-r}{y} p^r (1-p)^y \quad (4.11)$$

Utilizando essa distribuição o valor esperado de Y é dado pela equação:

$$E(Y) = r \frac{(1-p)}{p} \quad (4.12)$$

e a variância [28]:

$$Var(Y) = \frac{r(1-p)}{p^2} \quad (4.13)$$

Se definimos a média como $\mu = \frac{r(1-p)}{p}$, conseqüentemente $E(Y) = \mu$, e a variância pode ser calculada como [28]:

$$Var(Y) = \mu + \frac{1}{r} \mu^2 \quad (4.14)$$

onde, é mostrado que a variância é uma função quadrática da média.

4.6.3 Modelagem estatística da expressão de genes

Esta etapa é importante pois é nela onde os cálculos dos parâmetros usados nos testes estatísticos são executados.

Os *reads* sequenciados são uma amostragem do “estado real” dos fragmentos e devido a isso é plausível esperar, em uma situação hipotética, que as contagens sejam ligeiramente diferentes até se a mesma amostra for sequenciada duas vezes. Isto se deve ao fato do número de *reads* que podem ser capturados pela plataforma de sequenciamento ser finito, e por isso somente é possível obter uma amostragem do estado real das fontes que os geraram. Dado, então, que os experimentos de sequenciamento são considerados como uma amostragem randômica produzida pelos *reads* a partir de um *pool* de fragmentos, a representação natural das contagens pode ser a distribuição de Poisson, que está associada à processos de contagens (número de ocorrências de um evento por um certo período de tempo, por exemplo):

$$f(n, \lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \quad (4.15)$$

onde n é o número da contagem dos *reads* associado ao um gene e λ é o valor esperado

do número total de *reads* gerados pelos fragmentos dos transcritos que alinhariam ao gene. A característica de que até mesmo sequenciando duas amostras diferentes as contagens são diferentes, pode ser entendida como um ruído técnico, que é conhecida como *shot noise*, e esta variabilidade frequentemente pode ser bem associada ao ruído de Poisson em replicatas técnicas. Contudo, quando amostras são coletadas a partir de fontes biológicas distintas, a variância nas contagens associada a um gene na maioria das vezes é maior que a média, e esta situação proíbe o uso da distribuição de Poisson que é apropriada quando a variância é igual a média. Neste último caso é apropriado o uso da distribuição binomial negativa, que também pode substituir a distribuição de Poisson no acaso anterior [17], pois nesta a variância é maior que a média e é calculada da seguinte forma:

$$\nu = \mu + \alpha\mu^2 \quad (4.16)$$

onde α é o fator de dispersão e μ é a média [15, 19].

A estimativa do fator α é uma das diferenças fundamentais entre os pacotes edgeR e DESeq. Tal estimativa no edgeR é calculada a partir da combinação ponderada de dois componentes: um efeito da dispersão específica para cada gene e um efeito de dispersão comum que afeta todos os genes. O cálculo da estimativa conduzido pelo DESeq separa a estimativa da variância em uma parte que acomoda os cálculos ligado à estimativa da expressão média do gene, e a uma segunda parte que é destinada à modelagem de um termo associado à variabilidade da expressão biológica [15, 19].

Um outro pacote que será analisado neste trabalho, Cuffdiff, possui dois tipos de cálculos de variância: uma associada a genes que possuem uma única isoforma e outra que está relacionada a genes que possuem múltiplas isoformas. Quando o gene possui uma única isoforma a variância é calculada analogamente ao DESeq e quando houver múltiplas isoformas o cálculo é feito a partir de um modelo que mistura modelos binomiais negativos usando parâmetros da distribuição beta como pesos. O pacote baySeq utiliza uma abordagem bayseana na modelagem de distribuições binomiais negativas, onde os parâmetros da probabilidade *a priori* são estimados por amostragem numérica a partir dos dados. O pacote PoissonSeq modela as contagens dos genes N_{ij} como uma variável de Poisson, onde a média μ_{ij} da distribuição é representada por uma relação log linear $\log \mu_{ij} = \log d_j + \log \beta_i + \gamma_i \delta_j$, onde d_j representa o tamanho da biblioteca normalizada, β_i é o nível de expressão do gene i e γ_i é a correlação do gene i com a condição δ_j e se não houver diferença significativa na expressão do gene entre duas condições então γ_i é zero. Por fim, o pacote limma inicialmente idealizado para análise de expressão diferencial foi atualizado de modo a ser possível a análise de dados vindos de RNA-seq, incorporando um método de normalização apropriado para então inferir expressão diferencial a partir de modelos lineares [15, 19].

Tabela 4.2: Tabela comparativa dos métodos de modelagem estatística da expressão de genes executados pelos pacotes que serão estudados juntamente com as informações do DESuvrel.

Pacote	Método
baySeq	Utiliza uma abordagem bayseana na modelagem de distribuições binomiais negativas.
edgeR	Assume que os <i>reads</i> seguem uma distribuição binomial negativa.
DESeq	Assume que os <i>reads</i> seguem uma distribuição binomial negativa.
CuffDiff	Quando o gene possuir uma única isoforma é assumido que os genes seguem uma distribuição binomial negativa e no caso de haver múltiplas isoformas é usado uma distribuição binomial negativa com parâmetros da distribuição beta como pesos.
PoissonSeq	Modela as contagens dos genes N_{ij} como uma variável de Poisson.
limmaQN	Utiliza modelos lineares.
limmaVoom	Utiliza modelos lineares.
DESuvrel	As relevâncias dos genes são obtidas a partir das contagens sem nenhum tipo de suposição.

4.7 Teste para expressão diferencial

Nesta etapa de um experimento RNA-Seq é onde propriamente a inferência de expressão diferencial é executada, a partir da estimativa dos parâmetros dos respectivos modelos estatísticos entre dependendo do pacote, duas condições ou mais [15]. Para executar esta tarefa, os pacotes edgeR e DESeq usam uma variação do teste exato de Fisher adequado à distribuição binomial negativa para calcular valores-p. O pacote limma usa um teste t moderado com erros padrões e graus de liberdades modificados para calcular o valor-p [15]. O pacote Cuffdiff assume que y , a taxa de contagens normalizadas entre duas condições, aproximadamente segue uma distribuição normal e então usa um teste t para calcular o p-valor. O erro padrão é moderado entre todos os genes com um fator *shrinkage* que efetivamente atua com a intenção de usar informações de todos os genes para melhor a inferência de um único gene e os graus de liberdade são calculados a partir de um termo que a representa o número de graus de liberdade a priori [15]. O pacote baySeq utiliza uma abordagem bayseana para inferência de expressão diferencial, onde então, no algoritmo inicialmente é estimado dois modelos para todos gene, um assumindo que não ocorre expressão diferencial e o outro assumindo que existe e em seguida a partir dos dados a função verossimilhança adequada ao modelo é usada para identificar os genes expressos diferencialmente [15]. A abordagem utilizada no método implementado no pacote PoissonSeq realiza um teste de significância do termo γ_i , que é a correlação da expressão do gene i entre duas condições, calculado a partir de *scores* estatísticos [15]. Por fim, é importante lembrar que todos os paco-

Tabela 4.3: Tabela comparativa dos métodos de normalização executados pelos pacotes que serão estudados juntamente com as informações do DESuvrel.

Pacote	Método
baySeq	No algoritmo inicialmente é estimado dois modelos para todos os genes: um assumindo que não ocorre expressão diferencial e outro assumindo que existe, e em seguida a partir dos dados a função verossimilhança adequada ao modelo é usada para identificar os genes expressos diferencialmente.
edgeR	Varição do teste exato de Fisher.
DESeq	Varição do teste exato de Fisher.
CuffDiff	É assumido que a taxa de contagens normalizadas entre duas condições segue aproximadamente uma distribuição normal e então usa teste o t para calcular o p-valor.
PoissonSeq	Realiza um teste de significância do termo γ_i , que é a correlação da expressão do gene i entre duas condições, calculado a partir de <i>scores</i> estatísticos [15].
limmaQN	Utiliza um teste t moderado com erros padrões e graus de liberdades modificados para calcular o valor-p [15].
limmaVoom	Utiliza um teste t moderado com erros padrões e graus de liberdades modificados para calcular o valor-p [15].
DESuvrel	As relevâncias dos genes são obtidas a partir das contagens sem nenhum tipo de suposição.

tes possuem implementados em seus respectivos algoritmos, abordagens padrões para correção de hipótese múltipla com exceção do PoissonSeq que implementa uma nova forma de calcular a estimativa da taxa de falsa descoberta [15].

Capítulo 5

Metodologia

A inferência de expressão diferencial geralmente é feita por *softwares* como, por exemplo, o edgeR, associados à plataforma estatística R. Para a análise de desempenho da inferência de expressão diferencial entre os pacotes DESeq, edgeR, PoissonSeq, Cuffdiff, limma baySeq e a implementação do método Suvrel, chamada de DESuvrel, inicialmente será descrita a metodologia utilizada para realizar esta análise. Após essa descrição, na seção seguinte será descrito como aplicar o método Suvrel na inferência e, por fim, será descrita as características do conjunto de dados que será usado na análise da performance.

5.1 Curvas *ROC*

As primeiras utilizações das curvas ROC datam da Segunda Guerra Mundial, sendo usadas na detecção de sinais eletrônicos e problemas com radares. Naquela época, tais curvas eram usadas para quantificar a habilidade (*Receiver Operating Characteristic, ROC*) dos operadores de radar (*receiver operators*) em distinguir um sinal de um ruído. Essa medida foi importante pois estava associada a habilidade de um operador decidir corretamente se um sinal no radar era um avião inimigo (sinal) ou algum outro objetivo irrelevante (ruído). Nos anos seguintes, as curvas *ROC* foram utilizadas em psicologia experimental e, posteriormente nos anos 70, foram largamente usadas na classificação de pessoas doentes ou não na área de pesquisa biomédica. Já no da década de 90, as curvas começaram a ser adotadas no aprendizado de máquina (*machine learning*) e foi a partir deste estudos que ficou demonstrada a importância delas na avaliação e comparação de algoritmos [29].

Ao se estudar uma instância (gene) de um assunto (expressão diferencial),

cada instância pode ser positiva p (expresso diferencialmente) ou negativa n (não expresso diferencialmente) gerando a classe p, n , e com ajuda de um método de inferência elas podem ser classificadas como positiva Y ou negativa N , gerando a classe preditiva Y, N . O mapeamento preditivo de cada instância pode ser produzido por um classificador a partir de valores contínuos e a predição de qual classe o elemento pertencerá será feita usando um limiar, ou o classificador pode ser discreto predizendo a classe do elemento [30].

Com as últimas definições, é possível aprofundar o estudo deste tópico a partir da necessidade hipotética de predizer duas classes. Dado um classificador e uma instância, existem quatro desfechos: a instância é positiva e é classificada como positiva, assim é denominada como um verdadeiro positivo; se a instância é positiva e é classificada como negativa, ela é denominada como um falso negativo; se instância é negativa e é classificada como negativa, ela é então denominada como um verdadeiro negativo e se for classificada como positiva, ela é denominada como falso positivo. Logo, podemos utilizar essas informações para determinar algumas métricas que serão usadas para a construção das futuras curvas ROC. A seguir seguem tais definições [30].

Taxa de verdadeiro positivo:

$$taxa_{vp} = \frac{\text{positivos corretamente classificados}}{\text{total de positivos}} \quad (5.1)$$

Taxa de falso positivo:

$$taxa_{fp} = \frac{\text{negativos incorretamente classificados}}{\text{total de negativos}} \quad (5.2)$$

Sensitividade = $taxa_{vp}$

e por fim

$$Especificidade = \frac{\text{verdadeiros negativos}}{\text{falsos positivos} + \text{verdadeiros negativos}} \quad (5.3)$$

A partir deste ponto é possível definir um espaço bidimensional ROC como sendo constituído pela taxa de verdadeiro positivo no eixo Y e a taxa de falso positivo sendo o eixo X. Quando um classificador discreto é aplicado a um conjunto teste, ele produz somente uma classe de decisão Y ou N , gerando um ponto no espaço ROC. Porém, um outro classificador probabilístico pode gerar uma probabilidade ou *score*, indicando um valor numérico que representa o grau ao qual a instância é membro de uma classe. É importante notar que existe diferença entre classificadores que geram

Tabela 5.1: Tabela de *scores* de um classificador probabilístico hipotético, onde p e n representam uma classe positiva e negativa, respectivamente

Instância	Classe	Score
1	p	0,9
2	p	0,8
3	n	0,7
4	p	0,6
5	p	0,55
6	p	0,54
7	n	0,53
8	n	0,52
9	p	0,51
10	n	0,505
11	p	0,4
12	n	0,39
13	p	0,38
14	n	0,37
15	n	0,36
16	n	0,35
17	p	0,34
18	n	0,33
19	p	0,30
20	n	0,1

probabilidades e os que geram *scores*, embora possuam o mesmo nome. Um classificador gera probabilidade se responder estritamente aos teoremas padrões da probabilidade ao passo que se o saída do classificador não responder, ele então gera um *score*, onde este último exemplo somente indicará que um *score* mais alto corresponderá a uma probabilidade mais alta [30].

A tabela 5.1, adaptada da referência [30], mostra os resultados de um classificador probabilístico hipotético, onde a primeira coluna **Instância** identifica a instância, a coluna seguinte, nomeada de **Classe**, indica qual classe originalmente a instância pertence (p e n representa positiva e negativa, respectivamente) e a última coluna, **Score**, indica o *score* associado à classificação desta instância. Uma curva ROC como a da Figura 5.1 gerada a partir dos dados da Tabela 5.1 (onde por necessidade do algoritmo, os *scores* devem estar ordenados de forma decrescente) se inicia considerando um nível auxiliar de $+\infty$ gerando o ponto (0;0). Este nível é decrescido até encontrar a primeira instância positiva com *score* 0,9 gerando o ponto (0;0,1). Em um espaço bidimensional ROC, um ponto é definido como (*taxa de falso positivo*; *taxa de verdadeiro positivo*) e dado que a primeira instância é positiva, então o valor 0,1 é atribuído ao eixo da taxa de verdadeiro positivo e o seu valor é obtido com o auxílio

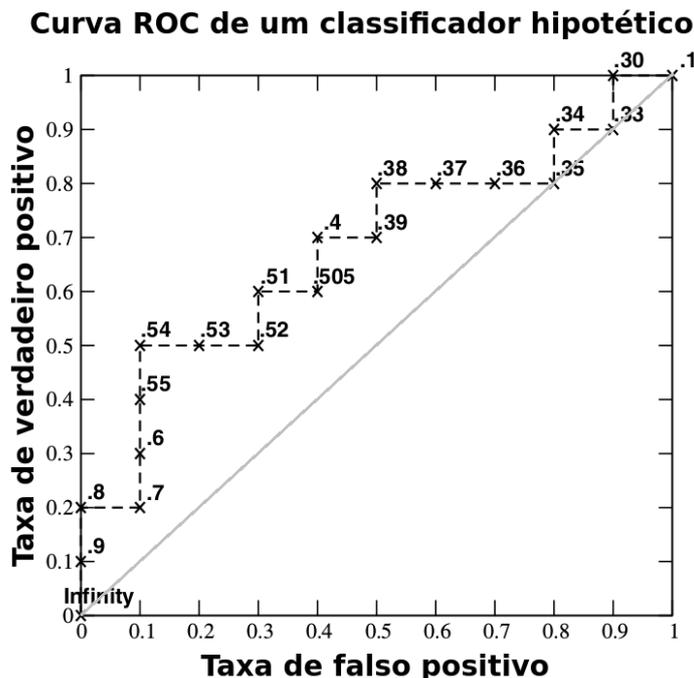


Figura 5.1: Curva ROC gerada a partir da tabela 5.1. Fonte: [30].

da equação 5.1 a partir do seguinte cálculo [30]:

$$taxa_vp = \frac{1}{10} = 0,1 \quad (5.4)$$

O nível auxiliar é cada vez mais reduzido e a ideia por trás da construção da curva continua a mesma até todas as instâncias terem sido analisadas quando o nível chega a 0,1 e o ponto (1;1) é plotado. É importante notar que qualquer curva ROC é gerada a partir de um conjunto finito de instâncias e é, na verdade, uma função degrau, onde ela se aproxima de uma curva verdadeira a medida que o número de instâncias se aproxima do infinito. Por fim, no gráfico acima é possível visualizar uma linha auxiliar que representa a curva que um classificador aleatório geraria [30].

Analisar visualmente a performance de vários classificadores por meio de curvas pode se tornar subjetivo e, por consequência, é gerada a necessidade de lançar mão de um método que reduza a curva à um valor escalar. Tal tarefa é possível por meio do cálculo da área abaixo da curva ROC (ou abreviadamente AAC). Desde que o valor é gerado a partir do cálculo de uma porção da área de um quadrado unitário, seu valor sempre estará entre 0 e 1, sendo que um classificador randômico deve possuir um valor de 0,5 e um classificador perfeito gerará um valor de 1 [30].

5.2 R estatística

A análise estatística de dados provenientes de RNA-seq pode ser feita dentro do ambiente R estatística [31]. O R é uma linguagem de programação e um ambiente de desenvolvimento integrado utilizado no ambiente acadêmico para cálculos estatísticos e processamento de dados. Seu código fonte é livre e permite a chamada de funções específicas de bibliotecas como os pacotes da área de análise de genes expressos diferencialmente, que serão escritos na seção 5.3.

5.3 Pacotes utilizados neste estudo para inferência de expressão diferencial

Nas seções anteriores foram descritos os passos envolvidos na inferência de expressão diferencial, que se inicia com o sequenciamento das amostras, onde elas são mapeadas e contadas e, por fim, a análise da expressão é executada. O ranqueamento dos genes envolvidos em expressão diferencial é feito majoritariamente no ambiente estatístico R utilizando ferramentas públicas, sendo que entre elas existem algumas que continuam sendo editadas e outras que foram substituídas por uma nova compilação. Entre os pacotes disponíveis, este estudo analisará os pacotes edgeR, DESeq, baySeq, PoissonSeq, limmaQN (limma utilizando a normalização por quantil), limma-Voom (limma utilizando a normalização pela função *voom*) e CuffDiff na comparação de performance da implementação do método Suvrel, chamada de DESuvrel.

5.4 Uso do método Suvrel na inferência de expressão diferencial em dados originados de RNA-Seq

Os passos para obtenção das relevâncias de cada gene seguem a referência [14], que origina o Método Suvrel, os quais são estruturados a partir da definição de uma métrica seguida pela definição de uma função custo, tornando possível o cálculo das relevâncias. Afim de iniciar a metodologia, inicialmente é designado o vetor $\{x_{ik}\}$, que representará a contagem do gene k no experimento i , seguida pela definição da distância D_{ij}^2 entre dois experimentos i e j

$$D_{ij}^2 = \sum_k \omega_k \|x_{ik} - x_{jk}\| \quad (5.5)$$

Em muitos estudos envolvendo experimentos associados a um conjunto de parâmetros, tem-se o objetivo de encontrar os parâmetros mais relevantes para diferenciar tais experimentos, assim como tem-se o objetivo de encontrar os genes mais relevantes envolvidos em expressão diferencial a partir de um conjunto de experimentos. Logo, a inserção da variável ω_k no cálculo da distância, que ainda necessita ser definida, está associada a esta constatação e é ela que nos fornecerá o grau de relevância de um determinado gene.

Assim, para continuar o desenvolvimento de uma metodologia que forneça os cálculos das relevâncias, define-se uma função objetivo:

$$E = \frac{1}{n_{AA}} \sum_{ij \in A} D_{ij}^2 + \frac{1}{n_{BB}} \sum_{ij \in B} D_{ij}^2 - \frac{\gamma}{n_{AB}} \sum_{i \in A, j \in B} D_{ij}^2 \quad (5.6)$$

Na equação 5.6, o primeiro termo calcula o somatório das distâncias entre os experimentos de uma mesma condição A dividido pelo número de pares pertencentes à condição A ; o segundo termo é semelhante ao primeiro diferindo somente quanto ao cálculo que será feito nos experimentos que fazem parte da condição B ; e o terceiro termo fornece o cálculo do somatório das distâncias entre os experimentos das condições A e B dividido pelo número de pares que é possível produzir entre os experimentos das duas condições. Por fim, γ é parâmetro que irá controlar a importância dada ao termo que calcula as distâncias entre condições diferentes. A equação 5.6 pode ser reescrita como:

$$E = \sum_k \omega_k \left(\frac{1}{n_{AA}} \sum_{ij \in A} \|x_{ik} - x_{jk}\| + \frac{1}{n_{BB}} \sum_{ij \in B} \|x_{ik} - x_{jk}\| - \frac{\gamma}{n_{AB}} \sum_{i \in A, j \in B} \|x_{ik} - x_{jk}\| \right) \quad (5.7)$$

$$E = \sum_k \omega_k \epsilon_k \quad (5.8)$$

onde,

$$\epsilon_k = \frac{1}{n_{AA}} \sum_{ij \in A} \|x_{ik} - x_{jk}\| + \frac{1}{n_{BB}} \sum_{ij \in B} \|x_{ik} - x_{jk}\| - \frac{\gamma}{n_{AB}} \sum_{i \in A, j \in B} \|x_{ik} - x_{jk}\| \quad (5.9)$$

A observação de que quando a equação 5.6 for extrema negativamente os experimentos são próximos entre a mesma condição e distantes entre condições diferentes, juntamente com o vínculo $\sum_k \omega_k^2 = 1$, podem ser usado associados aos multiplicadores de Lagrange (com mais detalhes no apêndice A) para obter o cálculo da relevância de cada gene:

$$\omega_k = \frac{-\epsilon_k}{\sqrt{\sum_k \epsilon_k^2}} \quad (5.10)$$

A equação para o cálculo da relevância 5.10 foi obtida sem a definição

explícita da distância entre dois experimentos D_{ij} , o que torna a equação 5.10 também não explícita. Logo, para tornar o cálculo explícito para que seja possível a implementação em um código de programação, será utilizada a distância euclidiana, onde

$$D_{ij}^2 = \sum_k \omega_k (x_{ik} - x_{jk})^2 \quad (5.11)$$

Utilizando esta distância, a equação 5.12 pode ser calculada como (o apêndice B apresenta a dedução detalhada desta equação):

$$\epsilon_k = 2(\sigma_A^2 + \sigma_B^2) - \gamma[\sigma_A^2 + \sigma_B^2 + (\mu_A - \mu_B)^2] \quad (5.12)$$

e quando $\gamma = 1$, tem-se

$$\epsilon_k = \sigma_A^2 + \sigma_B^2 - (\mu_A - \mu_B)^2 \quad (5.13)$$

e quando $\gamma = 2$,

$$\epsilon_k = 2(\mu_A - \mu_B)^2 \quad (5.14)$$

Assim, uma das formas de calcular a relevância de cada gene k é a partir da seguinte expressão, utilizando as equações 5.10 e 5.14:

$$\omega_k = \frac{-(2(\mu_A - \mu_B)^2)}{\sqrt{\sum_k (2(\mu_A - \mu_B)^2)^2}} \quad (5.15)$$

onde μ_A é média das contagens da condição A atribuídas ao gene k e μ_B é a media das contagens na condição B e σ_A^2 é a variância das contagens da condição A . Embora seja apontado na literatura que métodos não-paramétricos não possuem performance superior aos paramétricos devido à pequena quantidade de experimentos [19], neste trabalho serão mostrados resultados que não seguem estas constatações, utilizando, por exemplo as equações 5.10 e 5.14 no calculo das relevâncias. No caso de haver mais do que duas condições, pode-se generalizar a metodologia da seguinte forma:

$$E = \sum_{l=1} \sum_{ij \in c_l} D_{ij}^2 - \gamma \sum_{l=1}^{n-1} \sum_{o=l+1}^n \left(\sum_{(i \in c_l), (j \in c_o)} D_{ij}^2 \right) \quad (5.16)$$

$$E = \sum_k \omega_k \left[\left(\sum_{l=1} \sum_{ij \in c_l} \|x_{ik} - x_{jk}\| \right) - \gamma \left(\sum_{l=1}^{n-1} \sum_{o=l+1}^n \frac{1}{n_{cl} n_{co}} \sum_{(i \in c_l), (j \in c_o)} \|x_{ik} - x_{jk}\| \right) \right] = \sum_k \omega_k \epsilon_k \quad (5.17)$$

E considerando a distância euclidiana ϵ_k torna-se:

$$\epsilon_k = 2 \sum_{l=1}^n \sigma_{cl}^2 - \gamma \left[\sum_{l=1}^{n-1} \sum_{o=l+1}^n (\mu_{cl} - \mu_{co})^2 + \sum_{l=1}^n \sigma_{cl}^2 \right] \quad (5.18)$$

e quando $\gamma = 1$, ϵ_k torna-se:

$$\epsilon_k = \sum_{l=1}^n \sigma_{cl}^2 - \sum_{l=1}^{n-1} \sum_{o=l+1}^n (\mu_{cl} - \mu_{co})^2 \quad (5.19)$$

e quando $\gamma = 2$

$$\epsilon_k = -2 \sum_{l=1}^{n-1} \sum_{o=l+1}^n (\mu_{cl} - \mu_{co})^2 \quad (5.20)$$

Nos cálculos anteriores considerou-se somente um gene de cada vez sem relacioná-lo com outros. Contudo o método Survrel pode ser expandido, analisando a relevância de um par de genes μ e ν , tornando possível o cálculo de todos os possíveis pares. Para obter a equação que nos forneça tais relevâncias, serão seguidos passos semelhantes aos descritos acima, também retirados da [14]. Assim, definindo a distância d_{ij}^2 entre dois experimentos i e j associando os genes μ e ν da seguinte maneira:

$$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (5.21)$$

A função custo pode ser escrita como:

$$E = \sum_{A \in C} \frac{1}{n_A} \sum_{i, j \in A} d_{ij}^2 - \gamma \sum_{\langle A \neq B \rangle \in C} \frac{1}{n_{AB}} \sum_{i \in A, j \in B} d_{ij}^2 \quad (5.22)$$

onde A e B referem-se às condições, C denota as condições experimentais, γ é um parâmetro que irá controlar a importância dada ao termo que calcula as distâncias entre condições diferentes, $\langle A \neq B \rangle$ denota que os pares são incluídos somente uma única vez; n_A é o número de elementos na condição A , $n_{AB} = n_A n_B$ (números de elementos da condição A multiplicado pelo número de elementos na condição B) e $\sum_{i \in A}$ é o somatório sobre todas as amostras que pertencem a condição A . Como pode ser visto com mais detalhes no apêndice C a descrição matemática de todos os passos seguintes, a função custo pode ser definida como:

$$E = \sum_{\mu, \nu} g_{\mu\nu} \epsilon_{\mu\nu} \quad (5.23)$$

onde,

$$\epsilon_{\mu\nu} = e_{\mu\nu}^{in} - \gamma e_{\mu\nu}^{out} \quad (5.24)$$

$$\epsilon_{\mu\nu}^{in} = \sum_{a \in C} \frac{1}{n_{aa}} \sum_{i, j \in a} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (5.25)$$

Tabela 5.2: Exemplo de um tensor métrico hipotético produzido após três genes 1, 2 e 3 terem sido alvos da aplicação do método Suvrel.

	Gene 1	Gene 2	Gene 3
Gene 1	g_{11}	g_{12}	g_{13}
Gene 2	g_{21}	g_{22}	g_{23}
Gene 3	g_{31}	g_{32}	g_{33}

$$\epsilon_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} \frac{1}{n_{ab}} \sum_{(i \in a), j \in b} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (5.26)$$

e

$$\epsilon_{\mu\nu} = 2 \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma(K-1) \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) + K^2 cov(m_{\mu}, m_{\nu}) \quad (5.27)$$

onde x_{μ}^a é o conjunto das contagens do gene μ que pertencem a condição A , $m_{\mu} = \{m_{a\mu}\}$ é o conjunto das médias das condições, sendo que $m_{a\mu} = \frac{1}{n_a} \sum_{i \in a} x_{i\mu}$ e K é o número de componentes de uma classe. $\epsilon_{\mu\nu}$ pode ser escrito como na Equação 5.28 e esta última equação será uma das formas que será estudada no capítulo 6.

$$\epsilon_{\mu\nu} = [2 - (K-1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu}) \quad (5.28)$$

De forma análoga a descrita no apêndice A, a observação de que quando a equação 5.22 for extrema negativamente os experimentos são próximos entre a mesma condição e distantes entre condições diferentes assim como o uso do vínculo $\sum_{\mu\nu} g_{\mu\nu}^2 = 1$, podem ser associados aos multiplicadores de Lagrange na obtenção da expressão para o cálculo do tensor métrico $g_{\mu\nu}$:

$$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum \mu' \nu' \epsilon_{\mu' \nu'}^2}} \quad (5.29)$$

calculado com o auxílio da equação 5.28. Se $\gamma^* < \frac{2}{K-1}$ e $\gamma > \gamma^*$, então o tensor métrico é definido positivamente [14]. Quando $K = 2$ (duas condições experimentais), $\gamma^* < 2$ e então $\gamma > 2$. Se $\gamma < 2$ o tensor métrico torna-se negativo (como já mencionado) e se $\gamma > 2$ o primeiro termo da equação 5.28 torna-se negativo, fazendo com que aumentos nas covariâncias intraclasses diminuam a função custo, indo em desencontro com as premissas iniciais do método. Dessa forma, um valor adequado para classificação é $\gamma = 2$ [14].

A utilização das Equações 5.28 e 5.29 produzirá um tensor métrico, como, por exemplo o representado na Tabela 5.2, que foi construído a partir de uma situação hipotética onde haviam somente 3 genes para serem analisados. A partir deste tensor

Tabela 5.3: Exemplo de um tensor métrico hipotético produzido após três genes 1, 2 e 3 terem sido alvos da aplicação do método Suvrel, usado para exemplificar como são obtidas as relevância usando a diagonal principal.

	Gene 1	Gene 2	Gene 3
Gene 1	g_{11}	g_{12}	g_{13}
Gene 2	g_{21}	g_{22}	g_{23}
Gene 3	g_{31}	g_{32}	g_{33}

Tabela 5.4: Exemplo de um tensor métrico hipotético produzido após três genes 1, 2 e 3 terem sido alvos da aplicação do método Suvrel para exemplificar como é obtida a relevância de cada gene atribuindo maior importância a ele mesmo através do primeiro termo, e somando um segundo termo que calcula a média dos pares que ele pode fazer.

	Gene 1	Gene 2	Gene 3
Gene 1	g_{11}	g_{12}	g_{13}
Gene 2	g_{21}	g_{22}	g_{23}
Gene 3	g_{31}	g_{32}	g_{33}

métrico é possível calcular três variações da relevância de cada gene. A primeira possibilidade é obtida quando é definido que a relevância de cada gene depende somente dele a partir da utilização da equação 5.30, e isso equivale a utilizar somente a diagonal principal do tensor métrico, como pode ser visualizado na Tabela 5.3.

$$w_{\mu} = g_{\mu\mu} \quad (5.30)$$

Utilizando a Equação 5.31, onde x_{iF} é o número total de genes estudados, é possível calcular a relevância de cada gene, atribuindo maior importância a ele mesmo através do primeiro termo, somando um segundo termo que calcula a média dos pares que ele pode fazer. Essa possibilidade é exemplificada na Tabela 5.4, onde, para definir a relevância do **Gene 1** é utilizado o valor g_{11} em azul, representando o primeiro termo, e os valores em verde representam as células que deverão ser somadas para o cálculo da média, representando o segundo termo.

$$w_{\mu} = g_{\mu\mu} + \frac{1}{1 - x_{iF}} \sum_{\langle \mu \neq \nu \rangle} g_{\mu\nu} \quad (5.31)$$

Por fim, utilizando a Equação 5.32 é possível calcular a relevância de um gene dependendo da média das associações em pares que ele que pode realizar. Um exemplo deste cálculo pode ser visualizado na Tabela 5.5, onde, para calcular a relevância do **Gene 1** toda a linha com os valores em vinho serão somados para posterior cálculo da média.

Tabela 5.5: Exemplo de um tensor métrico hipotético produzido após três genes 1, 2 e 3 terem sido alvos da aplicação do método Suvrel para exemplificar como é obtida a relevância de um gene dependendo da média das associações em pares que ele que pode realizar.

	Gene 1	Gene 2	Gene 3
Gene 1	g_{11}	g_{12}	g_{13}
Gene 2	g_{21}	g_{22}	g_{23}
Gene 3	g_{31}	g_{32}	g_{33}

$$w_{\mu} = \frac{1}{x_{iF}} \sum_{\nu} g_{\mu\nu} \quad (5.32)$$

5.5 Conjunto de dados usado para análise

Para avaliar a performance dos pacotes destinados a inferência de expressão diferencial e a implementação do método Suvrel (DESuvrel), foi utilizada a metodologia e o conjunto de dados disponíveis na referência [15]. A matriz de contagens foi produzida após o sequenciamento, mapeamento e contagem de amostras que são parte do Consórcio para Controle de Qualidade de Sequenciamento (*Sequencing Quality Control Consortium*, abreviado *SEQC*). Cada uma das amostras foi gerada a partir de uma mistura de fontes biológicas e um conjunto de RNAs em concentrações conhecidas (*spike-in*) do Consórcio de Controle de RNA Externo (*External RNA Control Consortium*, abreviado *ERCC*). A conjunto de replicatas que caracteriza o grupo A é composta por 5 replicatas técnicas constituídas de RNA de referência humano universal (*Universal Human Reference RNA*, *UHR*, que por sua vez é composto de RNA total de 10 linhagens de células humanas), juntamente com 2% de volume da mistura ERCC 1. Outras 5 replicatas, que compõem o grupo B, são constituídas de RNA de referência do cérebro humano (*Ambion's Human Brain Reference RNA*, *HBRR*) com 2% de volume da mistura ERCC 2.

Como as misturas ERCC 1 e 2 são constituídas de uma mistura de 92 oligonucleotídios poli-A divididos em quatro subconjuntos com quatro taxas de proporções 1/2, 2/3, 1 e 4 predefinidas, a definição de um oligonucleotídio verdadeiro negativo é determinada a partir do log da mudança dessa proporção igual a 0, e as outras proporções indicarão os verdadeiros positivos [15]. Para facilitar a organização da seção dos resultados (capítulo 6) este conjunto de dados será referenciado como “**conjunto de dados ERCC**”.

Um outro conjunto de dados que foi usado para análise da performance foi gerado pelo Projeto de Controle da Qualidade de Microarranjo (*MicroArray Quality*

Control, MAQC). Neste empreendimento, uma variedade de plataformas de microarranjos e tecnologias, como, por exemplo, qRT-PCR TaqMan, foram avaliadas, usando um conjunto de amostras de RNA humano do tecido cerebral e outro conjunto contendo uma mistura de RNA vindos de diferentes tipos de tecidos. As medidas TaqMan qRT-PCR deste projeto consistem em valores de abundância para um conjunto de 1000 RNAs a partir de 4 replicatas em cada uma das duas amostras [32] e assim como realizado no estudo apresentado na referência [15] uma tabela gerada com estas informações de abundância será utilizada para análise da performance das inferências. Para facilitar a organização dos resultados (6) este conjunto de dados será referenciado como **“conjunto de dados TaqMan”**.

Capítulo 6

Resultados e Discussão

A análise da performance da implementação do método Suvrel será comparada com as dos pacotes listados na Seção 5.3 no ranqueamento de genes expressos diferencialmente, a partir do conjunto de dados exposto na seção 5.5 utilizando curvas ROC descritas na seção 5.1.

O processo de produção das curvas ROC executado neste trabalho pode ser separado em três etapas. Na primeira etapa, com o auxílio dos *scripts* fornecidos na referência [15], os resultados das execuções dos pacotes foram carregados (exceto o DESuvrel). Na segunda etapa, a tabela de contagens, também disponível na referência [15], pode ser normalizada utilizando as equações 4.3 ou 4.5 como descrito na Seção 4.5. Em seguida, a implementação do método Suvrel é aplicada e uma lista relacionando o gene e sua respectiva relevância é produzida. Por fim, em uma terceira etapa é produzida a curva ROC, utilizando, a partir de cada pacote, os valores-p ajustados, levando-se em conta a taxa de falsa descoberta, juntamente com os valores de relevância calculados usando o DESuvrel.

Para facilitar a compreensão dos resultados, as curvas ROC serão apresentadas juntamente com as informações utilizadas para sua produção, exibindo qual a normalização e as equações utilizadas no cálculo das relevâncias. Embora não seja explicitado, todas as curvas serão produzidas usando o valor de $\gamma = 2, 0$.

Para iniciar a produção dos resultados, o primeiro conjunto de dados analisado foi o TaqMan, onde, utilizando ele, as curvas ROC nas Figuras 6.1 e 6.3 foram geradas, atribuindo os transcritos verdadeiros negativos como sendo aqueles que tiveram valores do \log_2 da mudança de expressão média menores do que 0,5, e o restante como sendo verdadeiros positivos. As Figuras 6.1 e 6.3 foram geradas usando um limiar de 0,5, mas no sentido de aprofundar a análise, as Figuras 6.2 e 6.4 mostram os valores das áreas abaixo das curvas ROC para limiares que vão de 0,5 a 2,0.

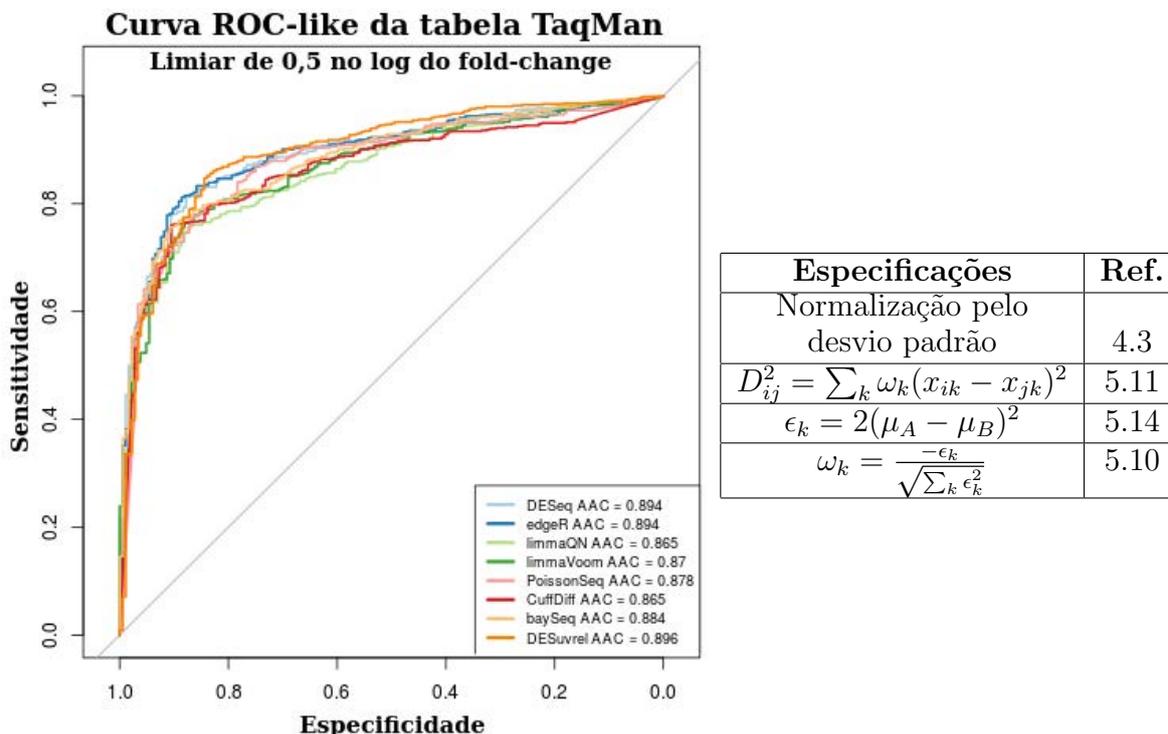


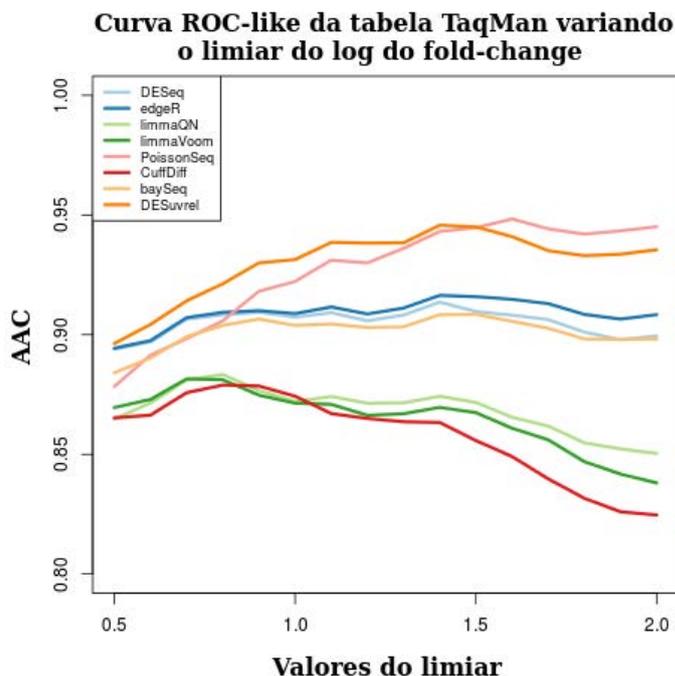
Figura 6.1: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem o \log_2 do *fold-change* maior que 0,5 são definidos como os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].

Visualizando as Figuras 6.1 e 6.3 é possível concluir que, embora todos os pacotes possuam valores próximos para a área abaixo da curva, a implementação do método Suvrel (chamada de DESuvrel) obteve a maior área. Com o auxílio das Figuras 6.2 e 6.4 é possível concluir que o DESuvrel possui as maiores áreas abaixo das curvas na maioria dos limiares analisados (aproximadamente 75%).

Em uma segunda etapa da produção dos resultados, foi utilizado o conjunto de dados ERCC por meio da tabela de dados que relaciona as misturas de 4 subgrupos de 92 oligonucleotídeos sintéticos nas taxas de 1/2, 2/3, 1 e 4 como descrito na Seção 5.5. Utilizando a proporção de $\log = 0$ (mistura de 1:1) para indicar quais oligonucleotídeos são os verdadeiros negativos e o restante sendo atribuídos como verdadeiros positivos, as curvas ROC nas Figuras 6.5 e 6.6 foram então geradas.

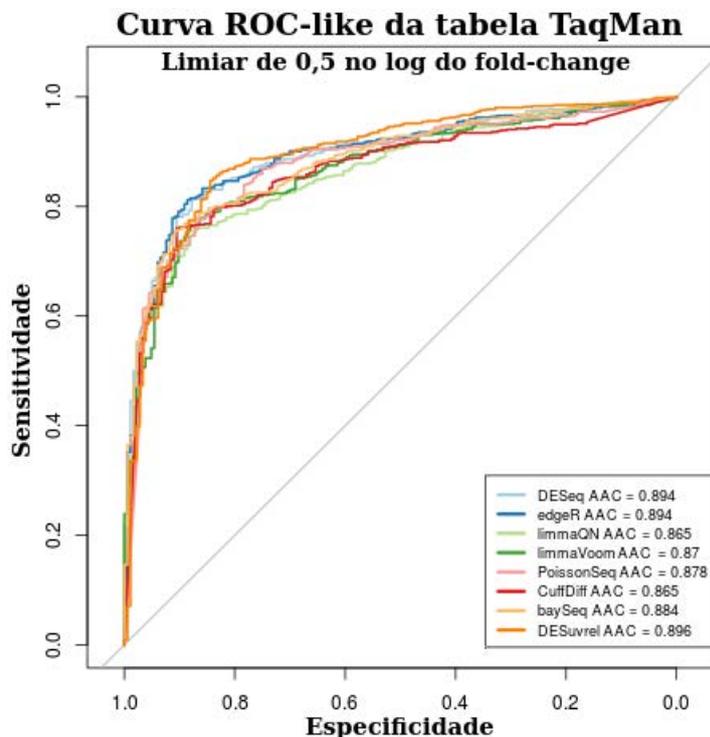
Visualizando as Figuras 6.5 e 6.6 é possível concluir que embora o DESuvrel, o pacote baySeq e o PoissonSeq possuam valores abaixo das curvas que destoam dos restantes, o DESuvrel obteve a maior.

Em uma terceira etapa foi analisada como a distância 5.21, que relaciona pares de genes, afeta o ranqueamento das relevâncias calculadas usando a diagonal principal do tensor métrico $g_{\mu\nu}$, através da equação 5.30. Os gráficos serão apresentados



Especificações	Ref.
Normalização pelo desvio padrão	4.3
$D_{ij}^2 = \sum_k \omega_k (x_{ik} - x_{jk})^2$	5.11
$\epsilon_k = 2(\mu_A - \mu_B)^2$	5.14
$\omega_k = \frac{-\epsilon_k}{\sqrt{\sum_k \epsilon_k^2}}$	5.10

Figura 6.2: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem \log_2 do *fold-change* maior que um limiar (0,5 a 2,0) são definidos como sendo os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].



Especificações	Ref.
Normalização média zero e variância 1	4.5
$D_{ij}^2 = \sum_k \omega_k (x_{ik} - x_{jk})^2$	5.11
$\epsilon_k = 2(\mu_A - \mu_B)^2$	5.14
$\omega_k = \frac{-\epsilon_k}{\sqrt{\sum_k \epsilon_k^2}}$	5.10

Figura 6.3: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem o \log_2 do *fold-change* maior que 0,5 são definidos como sendo os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].

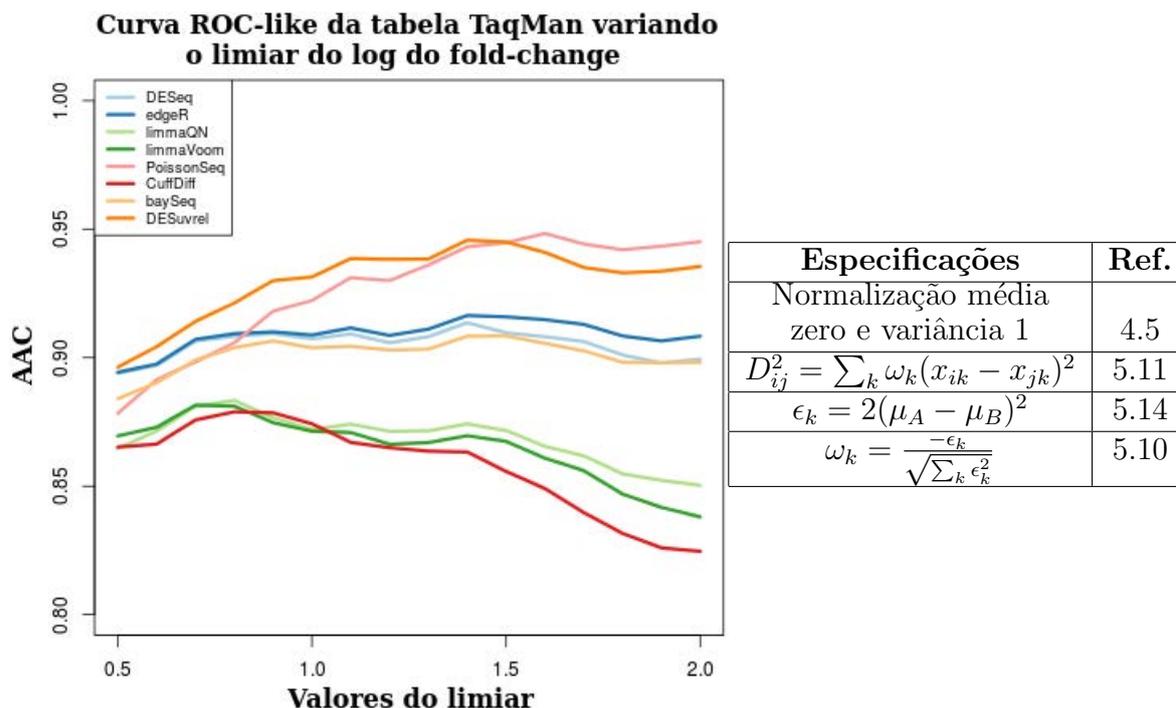


Figura 6.4: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem \log_2 do *fold-change* maior que um limiar (0,5 a 2,0) são definidos como sendo os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].

seguindo quase a mesma estrutura já apresentada até aqui. Na parte superior da figura será alocada a curva ROC e na parte inferior será alocada as informações que descrevem qual normalização e quais as equações utilizadas no cálculo da relevância. Embora não seja explicitado, todas as curvas serão produzidas usando o valor de $\gamma = 2.0$. Assim, dada a definição da situação que será analisada, o primeiro conjunto de dados analisado foi o TaqMan, onde, utilizando ele, a curva ROC na Figura 6.8 foi gerada, atribuindo os transcritos verdadeiros negativos como sendo aqueles que tiveram valores do \log_2 da mudança de expressão média menores do que 0,5, e o restante como sendo verdadeiros positivos. A Figura 6.8 foi gerada usando um limiar de 0,5, mas no sentido de aprofundar a análise, a Figura 6.9 mostra os valores das áreas abaixo das curvas ROC para limiares que vão de 0,5 a 2,0. Por fim, foi utilizado o conjunto de dados ERCC por meio da tabela de dados que relaciona as misturas de 4 subgrupos de 92 oligonucleotídeos sintéticos nas taxas de 1/2, 2/3, 1 e 4 como descrito na Seção 5.5. Utilizando a proporção de $\log = 0$ (mistura de 1:1) para indicar quais oligonucleotídeos são os verdadeiros negativos e o restante sendo atribuídos como verdadeiros positivos, a curva ROC na Figura 6.7 foi então gerada. Dado que utilizar as relevâncias quando a distância 5.11 é empregada equivale a utilizar as relevâncias dadas pela diagonal principal do tensor métrico $g_{\mu\nu}$ a partir da Equação 5.30, as conclusões delineadas

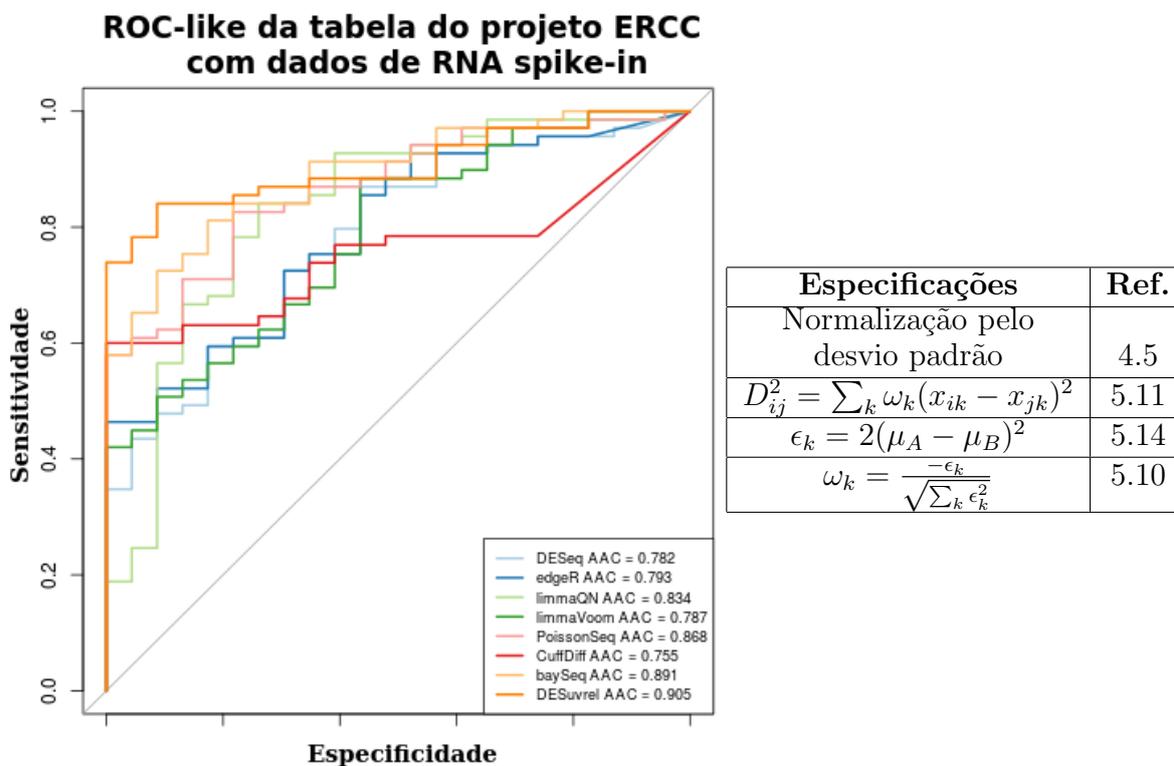


Figura 6.5: Análise ROC utilizando o conjunto de dados ERCC, onde o \log da proporção da mistura 0 é usado para indicar quais oligonucleotídios são os verdadeiros negativos, e o restante como sendo verdadeiros positivos e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].

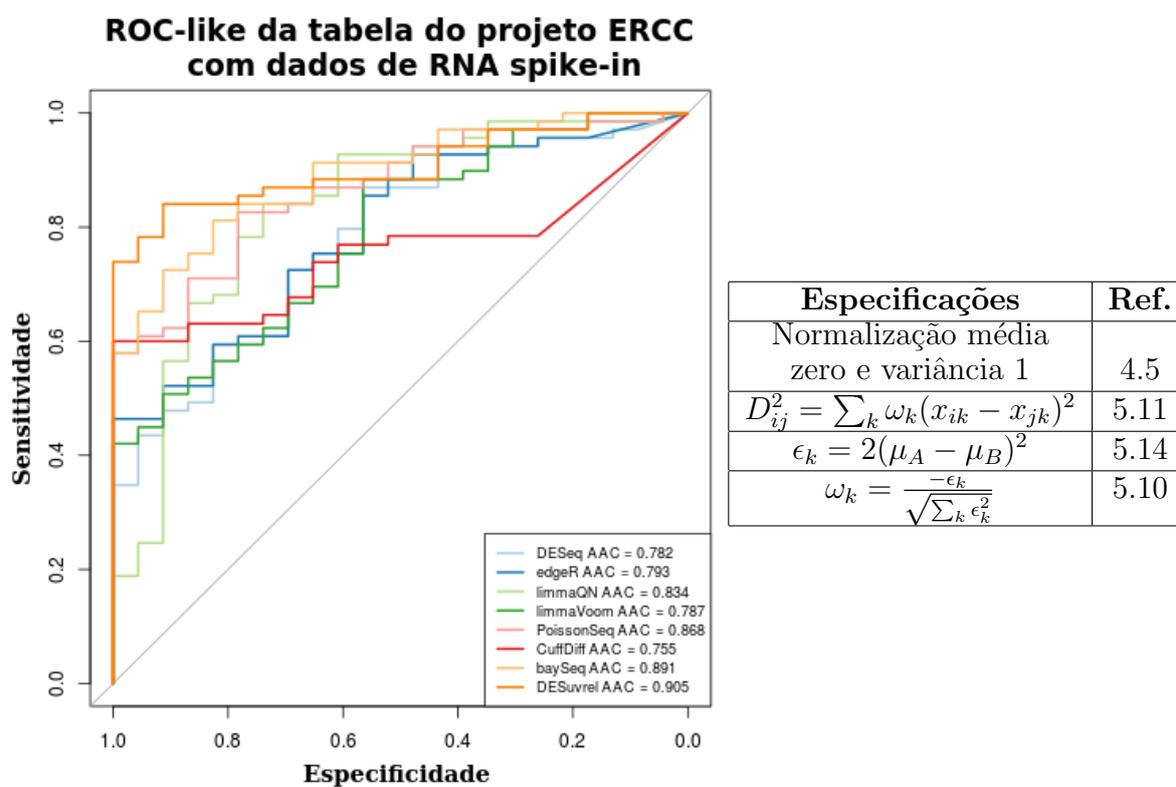
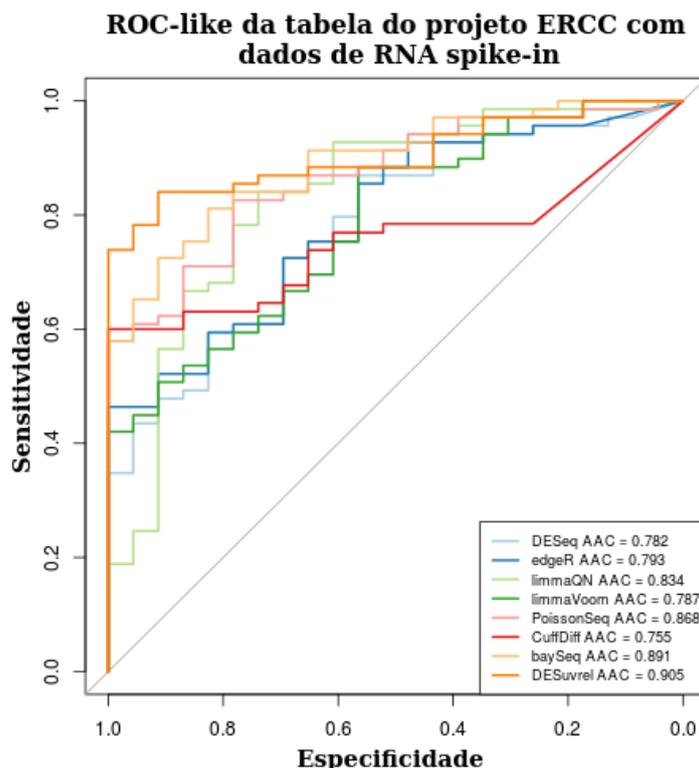


Figura 6.6: Análise ROC utilizando o conjunto de dados ERCC, onde o \log da proporção da mistura 0 é usado para indicar quais oligonucleotídios são os verdadeiros negativos, e o restante como sendo verdadeiros positivos e ao lado as especificações usadas na geração do gráfico. Adaptado de [15].

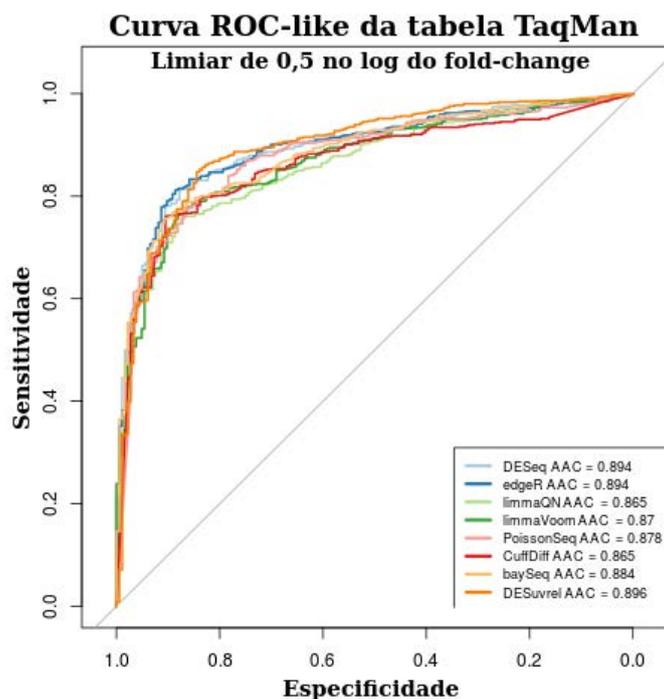


Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu_1 \nu_1} \epsilon_{\mu_1 \nu_1}^2}}$	5.29
$\omega_{\mu} = g_{\mu\mu}$	5.30

Figura 6.7: Análise ROC-like utilizando o conjunto de dados ERCC, onde o log da proporção igual 0 é usado para indicar quais oligonucleotídios são os verdadeiros negativos, e o restante como sendo verdadeiros positivos e abaixo as especificações usadas na geração do gráfico. Adaptado de [15].

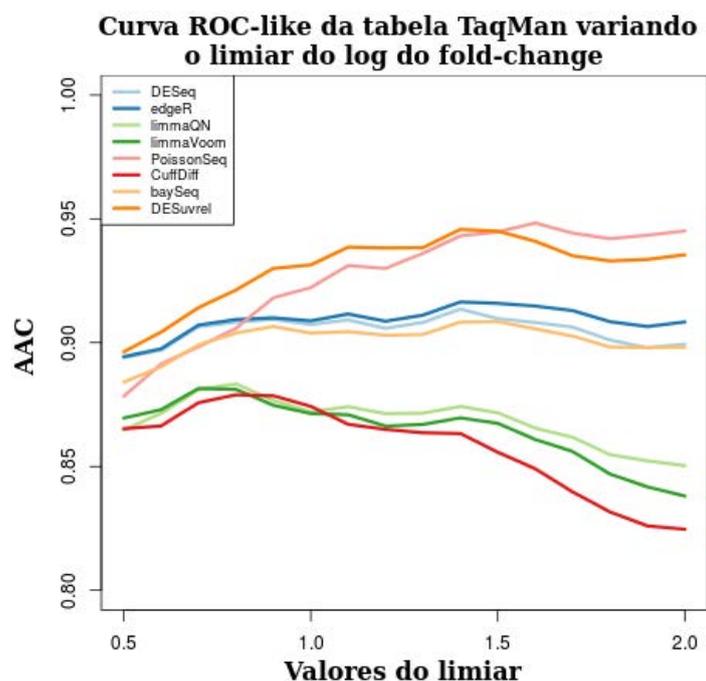
acima quanto ao desempenho do método Suvrel são exatamente as mesmas.

Em uma quarta etapa, caminhando no sentido de aprofundar a análise, foi estudado como é o comportamento das relevâncias calculadas utilizando as equações 5.30, 5.31, 5.32. Utilizando o conjunto de dados TaqMan, a curva ROC na Figura 6.10 foi gerada, atribuindo os transcritos verdadeiros negativos como sendo aqueles que tiveram valores do \log_2 da mudança de expressão média menores do que 0,5, e o restante como sendo verdadeiros positivos. Em seguida, foi utilizado o conjunto de dados ERCC por meio da tabela de dados que relaciona as misturas de 4 subgrupos de 92 oligonucleotídios sintéticos nas taxas de 1/2, 2/3, 1 e 4 como descrito na Seção 5.5. Utilizando a proporção de $\log = 0$ (mistura de 1:1) para indicar quais oligonucleotídios



Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu' \nu'} \epsilon_{\mu' \nu'}^2}}$	5.29
$\omega_{\mu} = g_{\mu\mu}$	5.30

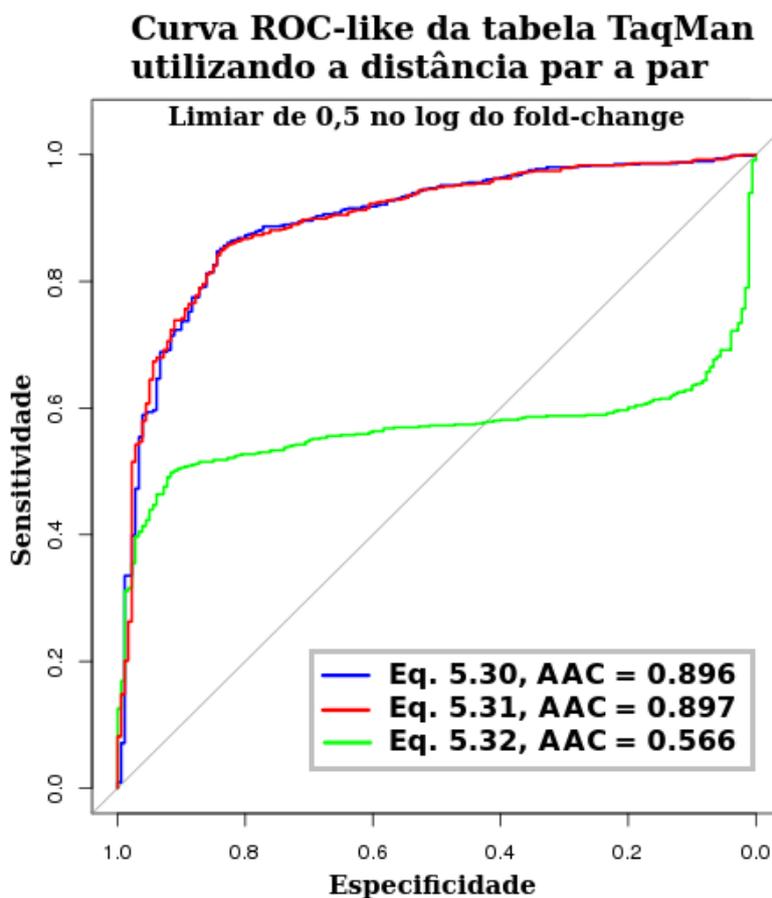
Figura 6.8: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem \log_2 do *fold-change* maior que 0,5 são definidos como sendo os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e abaixo as especificações usadas na geração do gráfico. Adaptado de [15].



Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu', \nu'} \epsilon_{\mu', \nu'}^2}}$	5.29
$\omega_{\mu} = g_{\mu\mu}$	5.30

Figura 6.9: Análise ROC-like a partir do conjunto de dados TaqMan, onde os transcritos que possuem o \log_2 do *fold-change* maior que um limiar (0,5 a 2,0) são definidos como sendo os verdadeiros positivos e os restantes como sendo os verdadeiros negativos, e abaixo as especificações usadas na geração do gráfico. Adaptado de [15].

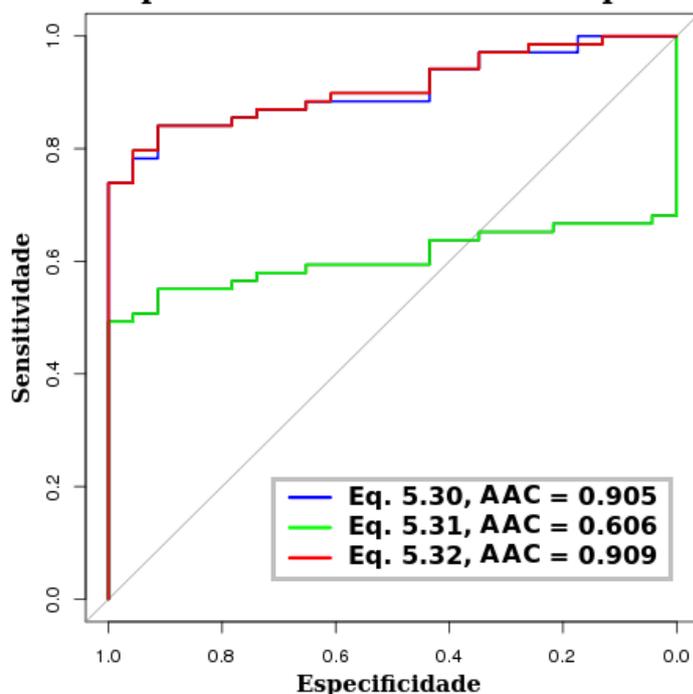
são os verdadeiros negativos e o restante sendo atribuídos como verdadeiros positivos, a curva ROC na Figura 6.11 foi então gerada. Visualizando as Figuras 6.10 e 6.11 é possível concluir que analisar a relevância de um transcrito empregando a associação que ele realiza com ele mesmo no mesmo patamar das associações de pares que ele pode fazer com outros transcritos, produz uma classificação aleatória, indicando que a relevância de cada transcrito é determinada predominantemente por ele mesmo.



Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu' \nu'} \epsilon_{\mu' \nu'}^2}}$	5.29
$w_{\mu} = g_{\mu\mu}$	5.30
$w_{\mu} = g_{\mu\mu} + \frac{1}{1-x_{iF}} \sum_{\langle \mu \neq \nu \rangle} g_{\mu\nu}$	5.31
$w_{\mu} = \frac{1}{x_{iF}} \sum_{\nu} g_{\mu\nu}$	5.32

Figura 6.10: Análise ROC a partir do conjunto de dados TaqMan, onde os transcritos que possuem \log_2 do *fold-change* maior que 0,5 são definidos como sendo os verdadeiros positivos e os restantes definidos como os verdadeiros negativos, e as relevâncias pelas equações 5.30, 5.31, 5.32. Abaixo as especificações usadas na geração do gráfico. Adaptado de [15].

ROC-like da tabela do projeto ERCC com dados de RNA spike-in utilizando distância par a par



Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu', \nu'} \epsilon_{\mu', \nu'}^2}}$	5.29
$w_{\mu} = g_{\mu\mu}$	5.30
$w_{\mu} = g_{\mu\mu} + \frac{1}{1-x_{iF}} \sum_{\langle \mu \neq \nu \rangle} g_{\mu\nu}$	5.31
$w_{\mu} = \frac{1}{x_{iF}} \sum_{\nu} g_{\mu\nu}$	5.32

Figura 6.11: Análise ROC utilizando o conjunto de dados ERCC, onde o log da proporção das misturas 0 é usado para indicar quais oligonucleotídios são os verdadeiros negativos, e o restante como sendo verdadeiros positivos e as relevâncias pelas equações 5.30, 5.31, 5.32. Abaixo as especificações usadas na geração do gráfico. Adaptado de [15].

Capítulo 7

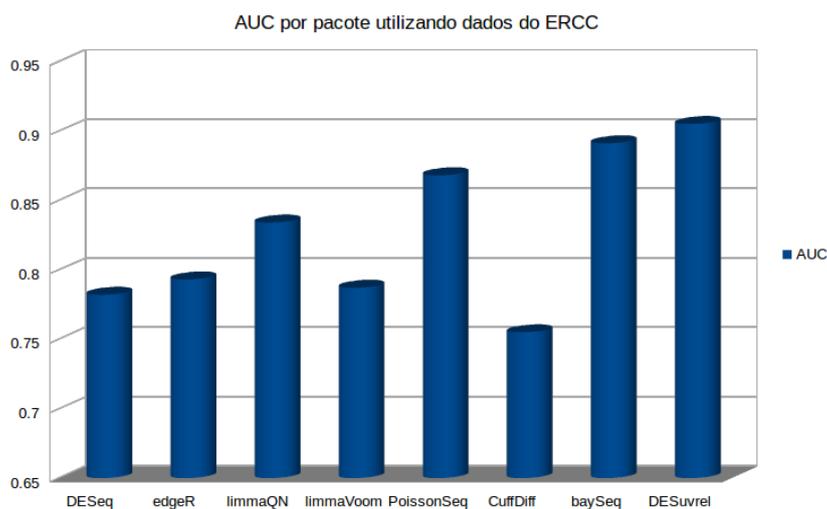
Conclusões e Perspectivas Futuras

Utilizando as informações associadas à análise ROC no Capítulo 6 por meio das Figuras 6.2, 6.4 e 6.9 e a Figura 7.1 a seguir, é possível concluir que as áreas abaixo das curvas do método Suvrel são maiores na grande maioria das situações analisadas.

Alinhada a conclusão anterior, é importante ressaltar que ela foi produzida a partir do método Suvrel, que não necessita de nenhuma suposição quanto a natureza da distribuição associada às contagens. O fato de não ser necessário nenhuma suposição deste tipo é o ideal, pois não é possível afirmar que as contagens seguem uma distribuição determinada em todas as situações que um pesquisador pode se deparar. Dessa forma, dado que o método proposto obteve as maiores áreas abaixo das curvas na grande maioria das situações, não faz nenhum tipo de suposição como mencionado e a sua execução é rápida, são pontos que podem ser usados para concluir que o DESuvrel está a frente dos pacotes.

Embora os resultados indiquem a viabilidade da aplicação do método Suvrel, é importante ressaltar que na literatura é apontado que métodos não-paramétricos teriam performance inferior aos métodos paramétricos [19], mas abordar este tipo de problema com o método Suvrel está mostrando que é necessário mais estudo neste tópico.

De uma forma geral, é importante ressaltar que o método pode ser usado na melhoria da performance de predição, como produzido na referência [14] em dados originados de microarranjos e concentração de metabólitos medidos por ressonância magnética, e o que foi desenvolvido neste estudo pode ser transportado e usado em outros problemas que envolvam ranqueamento de características, assim como também em predição.



Especificações	Ref.
Normalização média zero e variância 1	4.5
$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu})$	5.21
$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_{\mu}^a; x_{\nu}^a) - \gamma K^2 cov(m_{\mu}, m_{\nu})$	5.28
$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu, \nu} \epsilon_{\mu, \nu}^2}}$	5.29
$\omega_{\mu} = g_{\mu\mu}$	5.30

Figura 7.1: Gráfico em barras mostrando os valores abaixo da curva ROC, utilizando o conjunto de dados ERCC, onde o log da proporção igual 0 é usado para indicar quais oligonucleotídios são os verdadeiros negativos, e o restante como sendo verdadeiros positivos e abaixo as especificações usadas na geração do gráfico.

Referências Bibliográficas

- [1] Claire M. Fraser, Jonathan A. Eisen, Karen E. Nelson, Ian T. Paulsen, and Steven L. Salzberg. The value of complete microbial genome sequencing (you get what you pay for). *Journal of Bacteriology*, 184(23):6403–6405, 2002.
- [2] International Human Genome Sequencing Consortium Announces "Working Draft" of Human Genome. Disponível em: <<http://www.genome.gov/10001457>> , Acessado em: 22 de nov. de 2014.
- [3] Dez anos de genoma humano. Disponível em: <<http://cienciahoje.uol.com.br/colunas/deriva-genetica/dez-anos-de-genoma-humanog>> , Acessado em: 12 nov. 2014.
- [4] D. Husmeier, R. Dybowski, and S. Roberts. *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer-Verlag New York Inc, 2004.
- [5] T.D. Otto, M. Catanho, C. Tristão, M. Bezerra, R.M. Fernandes, G.S. Elias, A.C. Scaglia, B. Bovermann, V. Berstis, S. Lifschitz, A.B. Miranda, and W. Degrave. ProteinWorldDB: querying radical pairwise alignments among protein sets from complete genomes. *Bioinformatics*, 26(5):705–707, 2010.
- [6] J. Goecks, A. Nekrutenko, J. Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
- [7] J.A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10):671–682, 2011.
- [8] A. Oshlack, M.D. Robinson, and M.D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.
- [9] F. Ozsolak and P.M. Milos. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, 12(2):87–98, 2011.

-
- [10] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.
- [11] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 2015.
- [12] Charlotte Sonesson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(1):91, 2013.
- [13] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), January 2010. PMID: 19910308.
- [14] Boareto M.; Cesar J; Leite V.B.P.; Caticha N. Supervised variational relevance learning, an analytic geometric feature selection with applications to omic data sets. *Knowledge and Data Engineering, IEEE Transactions on Communications*, 2015.
- [15] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher Mason, Nicholas Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, 14(9):R95, 2013.
- [16] Ayat Hatem, Doruk Bozdog, Amanda Toland, and Umit Catalyurek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184, 2013.
- [17] Sonia Tarazona, Fernando García-Alcalde, Joaquin Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: A matter of depth. *Genome Research*, 2011.
- [18] Fatih Ozsolak and Patrice M. Milos. Single-molecule direct rna sequencing without cdna synthesis. *Wiley Interdisciplinary Reviews: RNA*, 2(4):565–570, 2011.
- [19] Francesca Finotello and Barbara Di Camillo. Measuring differential gene expression with rna-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, 2014.
- [20] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics*, 11(5):473–483, Sep 2010.

-
- [21] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, Apr 2010.
- [22] Thomas Hardcastle and Krystyna Kelly. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010.
- [23] Mark Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3):R25, 2010.
- [24] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- [25] Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 2011.
- [26] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.
- [27] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.
- [28] George Casella. *Statistical inference*. Thomson Learning, Australia Pacific Grove, CA, 2002.
- [29] Martinez E. Z. e Louzada Neto F. e Pereira B. B. A curva roc para testes diagnósticos. *Caderno de Saude Coletiva, Rio de Janeiro*, 11(1):7–31, 2003.
- [30] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.
- [31] The R Project for Statistical Computing. Disponível em: <<http://www.r-project.org>>, Acessado em: 22 de agos. 2012.
- [32] Bo Li and Colin Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

Apêndice A

Multiplicadores de Lagrange

O método dos multiplicadores de Lagrange fornece uma estratégia para encontrar o máximo ou o mínimo de uma função sujeita a certas restrições. Supõe-se que deseja-se minimizar uma função $f(x)$ em relação a uma variável x sujeito a restrição $g(x) = c$, no qual c é uma constante. A técnica dos multiplicadores de Lagrange consiste em definir uma função $\Lambda(x, \lambda)$ da seguinte maneira:

$$\Lambda(x, \lambda) = f(x) + \lambda[g(x) - c] \quad (\text{A.1})$$

com a constante λ chamada de multiplicador de Lagrange. Resolvendo $\nabla_{x,\lambda}\Lambda(x, \lambda) = 0$ obtém-se os valores de x que minimizam f . Utiliza-se esta abordagem para encontrar valores de ω_k que minimizam a função

$$E = \sum_k \omega_k \left[\frac{1}{n_c} \sum_{ij \in C} \|x_{ik} - x_{jk}\| - \frac{1}{n_f} \sum_{ij \in F} \|x_{ik} - x_{jk}\| \right] = \sum_k \omega_k \beta_k \quad (\text{A.2})$$

sujeito a

$$\sum_k \omega_k^2 = 1. \quad (\text{A.3})$$

Neste caso, define-se a função $\Lambda(\{\omega\}, \lambda)$ da seguinte maneira:

$$\Lambda(\{\omega\}, \lambda) = \sum_k \omega_k \beta_k + \lambda \left[\sum_k \omega_k^2 - 1 \right] \quad (\text{A.4})$$

O próximo passo é encontrar o mínimo desta função em relação à ω_k fazendo

$$\frac{\partial \Lambda(\{\omega\}, \lambda)}{\partial \omega_k} = 0 \quad (\text{A.5})$$

chegando em

$$\beta_k - 2\lambda\omega_k = 0 \tag{A.6}$$

Então,

$$\omega_k = \frac{\beta_k}{2\lambda} \tag{A.7}$$

Utilizando a restrição (A.3), encontra-se

$$\lambda = \frac{\pm\sqrt{\sum_k \omega_k^2}}{2}, \tag{A.8}$$

Uma vez que deseja-se encontrar o mínimo, escolhe-se o valor negativo. Logo,

$$\omega_k = \frac{-\beta_k}{\sqrt{\sum_k \omega_k^2}}. \tag{A.9}$$

Apêndice B

Obtenção da equação 5.12 utilizando a distância euclidiana

Utilizando a equação da distância euclidiana 5.11

$$D_{ij}^2 = \sum_k \omega_k (x_{ik} - x_{jk})^2 \quad (\text{B.1})$$

na função custo 5.6

$$E = \frac{1}{n_{AA}} \sum_{ij \in A} D_{ij}^2 + \frac{1}{n_{BB}} \sum_{ij \in B} D_{ij}^2 - \frac{1}{n_{AB}} \sum_{i \in A, i \in B} D_{ij}^2 \quad (\text{B.2})$$

tem-se a seguinte expressão:

$$E = \sum_k \omega_k \left(\frac{1}{n_A^2} \sum_{ij \in A} (x_{ik} - x_{jk})^2 + \frac{1}{n_B^2} \sum_{ij \in B} (x_{ik} - x_{jk})^2 - \frac{\gamma}{n_A n_B} \sum_{(i \in A), j \in B} (x_{ik} - x_{jk})^2 \right) \quad (\text{B.3})$$

$$E = \sum_k \omega_k \epsilon_k \quad (\text{B.4})$$

A variável ϵ_k pode ser escrita como:

$$\begin{aligned} \epsilon_k = & \frac{1}{n_A^2} \left(n_A \sum_{i \in A} x_{ik}^2 - 2 \sum_{i \in A} x_{ik} \sum_{j \in A} x_{jk} + n_A \sum_{j \in A} x_{jk}^2 \right) \\ & + \frac{1}{n_B^2} \left(n_B \sum_{i \in B} x_{ik}^2 - 2 \sum_{i \in B} x_{ik} \sum_{j \in B} x_{jk} + n_B \sum_{j \in B} x_{jk}^2 \right) \\ & - \frac{\gamma}{n_A^2 n_B^2} \left(n_B \sum_{i \in A} x_{ik}^2 - 2 \sum_{(i \in A)} x_i \sum_{(i \in A)} \sum_{(j \in A)} x_{jk} + n_A \sum_{(j \in B)} x_{jk}^2 \right) \end{aligned} \quad (\text{B.5})$$

$$\begin{aligned}
\epsilon_k &= \frac{1}{n_A} \sum_{i \in A} x_{ik}^2 - 2 \sum_{i \in A} \frac{x_{ik}}{n_A} \sum_{j \in A} \frac{x_{jk}}{n_B} + \frac{a}{n_A} \sum_{j \in A} x_{jk}^2 \\
&+ \frac{1}{n_B} \sum_{i \in B} x_{ik}^2 - 2 \sum_{i \in B} \frac{x_{ik}}{n_B} \sum_{j \in B} \frac{x_{jk}}{n_B} + \frac{1}{n_B} \sum_{j \in B} x_{jk}^2 \\
&- \gamma \sum_{i \in A} \frac{x_{ik}}{n_A} - \gamma \sum_{j \in B} \frac{x_{jk}^2}{n_B} - 2\gamma \sum_{i \in A} \frac{x_{ik}}{n_A} \sum_{j \in B} \frac{x_{jk}}{n_B}
\end{aligned} \tag{B.6}$$

Utilizando

$$\lambda^2 = \frac{1}{n} \sum_i x_i^2 - \left(\frac{1}{n} \sum_i x_i \right)^2 = \frac{1}{n} \sum_i x_i^2 - \mu^2 \tag{B.7}$$

$$\epsilon_k = 2(\sigma_A^2 + \sigma_B^2) - \gamma \sum_{i \in A} \frac{x_{ik}^2}{n_A} - \gamma \sum_{j \in B} \frac{x_{jk}^2}{n_B} + 2\gamma \mu_A \mu_B \tag{B.8}$$

E assim chega-se a equação 5.12

$$\epsilon_k = 2(\sigma_A^2 + \sigma_B^2) - \gamma[\sigma_A^2 + \sigma_B^2 + (\mu_A - \mu_B)^2] \tag{B.9}$$

Apêndice C

Obtenção da expressão da relevância usando a equação 5.21

Utilizando a equação da distância 5.21

$$d_{ij}^2 = \sum_{\mu, \nu} g_{\mu\nu} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (\text{C.1})$$

na função custo 5.22

$$E = \sum_{a \in C} \frac{1}{n_a} \sum_{i, j \in a} d_{ij}^2 - \gamma \sum_{\langle a \neq b \rangle \in C} \frac{1}{n_{ab}} \sum_{(i \in a), (j \in b)} d_{ij}^2 \quad (\text{C.2})$$

onde A e B referem-se às condições, C denotará as condições experimentais, γ é parâmetro que irá controlar a importância dada ao termo que calcula as distâncias entre condições diferentes, $\langle A \neq B \rangle$ denota que os pares são incluídos somente uma única vez; n_A é o número de elementos na condição A , $n_{AB} = n_A n_B$ (números de elementos da condição A multiplicado pelo número de elementos na condição B) e $\sum_{i \in a}$ é o somatório sobre todas as amostras que pertencem a condição A . A equação 5.22 pode ser escrita como:

$$E = \sum_{\mu, \nu} g_{\mu\nu} \epsilon_{\mu\nu} \quad (\text{C.3})$$

onde,

$$\epsilon_{\mu\nu} = e_{\mu\nu}^{in} - \gamma e_{\mu\nu}^{out} \quad (\text{C.4})$$

$e_{\mu\nu}^{in}$ pode ser escrita como:

$$e_{\mu\nu}^{in} = \sum_{a \in C} \frac{1}{n_{aa}} \sum_{i, j \in a} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (\text{C.5})$$

$$e_{\mu\nu}^{in} = \sum_{a \in C} \frac{1}{n_{aa}} \left[\sum_{i,j \in a} x_{i\mu} x_{i\nu} - \sum_{i,j \in a} x_{i\mu} x_{j\nu} - \sum_{i,j \in a} x_{j\mu} x_{i\nu} + \sum_{i,j \in a} x_{j\mu} x_{j\nu} \right] \quad (C.6)$$

$$e_{\mu\nu}^{in} = \sum_{a \in C} \frac{1}{n_a^2} \left[n_a \sum_{i \in a} x_{i\mu} x_{i\nu} + n_a \sum_{j \in a} x_{j\mu} x_{j\nu} - 2 \sum_{i,j \in a} x_{i\mu} x_{j\nu} \right] \quad (C.7)$$

$$e_{\mu\nu}^{in} = 2 \sum_{a \in C} \left[\frac{1}{n_a} \sum_{i \in a} x_{i\mu} x_{i\nu} - \frac{1}{n_a^2} \sum_{i,j \in a} x_{i\mu} x_{j\nu} \right] \quad (C.8)$$

$$e_{\mu\nu}^{in} = 2 \sum_{a \in C} cov(x_{i\mu}^a; x_{j\nu}^a) \quad (C.9)$$

onde,

$$cov(x_{i\mu}^a; x_{j\nu}^a) = \frac{1}{n_a} \sum_{i \in a} x_{i\mu} x_{i\nu} - \frac{1}{n_a^2} \sum_{i,j \in a} x_{i\mu} x_{j\nu} \quad (C.10)$$

$e_{\mu\nu}^{out}$ pode ser escrito como:

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} \frac{1}{n_{ab}} \sum_{(i \in a), (j \in b)} (x_{i\mu} - x_{j\mu})(x_{i\nu} - x_{j\nu}) \quad (C.11)$$

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} \frac{1}{n_{ab}} \left[\sum_{i \in a} x_{i\mu} x_{i\nu} - \sum_{(i \in a), (j \in b)} x_{i\mu} x_{j\nu} - \sum_{(i \in a), (j \in b)} x_{j\mu} x_{i\nu} + \sum_{j \in b} x_{j\mu} x_{j\nu} \right] \quad (C.12)$$

onde $n_{ab} = n_a n_b$

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} \frac{1}{n_a n_b} \left[n_b \sum_{i \in a} x_{i\mu} x_{i\nu} - \sum_{i \in a} x_{i\mu} \sum_{j \in b} x_{j\nu} - \sum_{i \in a} x_{i\nu} \sum_{j \in b} x_{j\mu} + n_a \sum_{j \in b} x_{j\mu} x_{j\nu} \right] \quad (C.13)$$

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} \left[\frac{1}{n_a} \sum_{i \in a} x_{i\mu} x_{i\nu} - \sum_{i \in a} \frac{x_{i\mu}}{n_a} \sum_{j \in b} \frac{x_{j\nu}}{n_b} - \sum_{i \in a} \frac{x_{i\nu}}{n_a} \sum_{j \in b} \frac{x_{j\mu}}{n_b} + \frac{a}{n_b} \sum_{j \in b} x_{j\mu} x_{j\nu} \right] \quad (C.14)$$

Utilizando

$$m_{a\mu} = \sum_{i \in a} \frac{x_{i\mu}}{n_a} \quad (C.15)$$

e

$$\frac{1}{n_a} \sum_{i \in a} x_{i\mu} x_{i\nu} = cov(x_{i\mu}^a; x_{i\nu}^a) + \sum_{i \in a} \frac{x_{i\mu}}{n_a} \sum_{j \in a} \frac{x_{j\nu}}{n_a} \quad (C.16)$$

vinda da covariância entre $x_{i\mu}^a$ e $x_{i\nu}^a$,

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} [cov(x_{i\mu}^a; x_{i\nu}^a) + m_{a\mu} m_{a\nu} - m_{a\mu} m_{b\nu} - m_{a\nu} m_{b\mu} + cov(x_{i\mu}^b; x_{i\nu}^b) + m_{b\mu} m_{b\nu}] \quad (C.17)$$

$$e_{\mu\nu}^{out} = \sum_{\langle a \neq b \rangle \in C} [cov(x_{i\mu}^a; x_{i\nu}^a) + cov(x_{i\mu}^b; x_{i\nu}^b) + (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu})] \quad (C.18)$$

$$(K - 1) = \sum_{a \in c} cov(x_\mu^a; x_\nu^a) + \sum_{\langle a \neq b \rangle \in C} (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu}) \quad (C.19)$$

onde K é o numero de experimentos de uma condição.

$$(K - 1) = \sum_{a \in C} cov(x_\mu^a; x_\nu^a) + K^2 cov(m_\mu; m_\nu) \quad (C.20)$$

onde $m_\mu = m_{a\mu}$ representa a média sobre uma condição experimental do gene μ . Enfim, $\epsilon_{\mu\nu}$ pode ser escrito como:

$$\epsilon_{\mu\nu} = [2 - (K - 1)\gamma] \sum_{a \in C} cov(x_\mu^a; x_\nu^a) - (\gamma)K^2 cov(m_\mu; m_\nu). \quad (C.21)$$

A equação C.22 que fornecerá a relevância de cada gene, pode ser obtida de maneira análoga a do apêndice A e pode ser calculada utilizando a expressão C.21.

$$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum \mu\nu' \epsilon_{\mu\nu'}^2}} \quad (C.22)$$

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, ____/____/____

Assinatura