UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE FÍSICA

Ciclo Celular Detalhado pela Análise de Componentes Principais

Lars Leonardo Sanhudo de Souza

Trabalho de Conclusão de Curso apresentado para a obtenção do grau de Bacharel no Curso de Física
Orientadora: Prof. Dra. Rita Maria Cunha de Almeida
Colaborador: Prof. Dr. Guido Lenz

Porto Alegre - RS

Dezembro de 2018

Agradecimentos

Muitas pessoas fizeram parte direta, ou indiretamente da realização deste trabalho. Primeiramente preciso agradecer a minha família, com todo o apoio emocional e material (sei que foi complicado, recompensarei a todos, um dia). Desde de 2013, com o sonho de entrar na UFRGS, de ainda fazer física, e a concretização desta primeira etapa, com este trabalho.

Também quero agradecer a todos os amigos que conheci da UFRGS em todos estes anos, em especial para o Gabriel e Juliana. Tivemos muitas discussões, diversão, e claro, momentos difíceis que enfrentamos juntos. Vocês são novos irmãos que conheci nesta vida.

Por fim, preciso agradecer a minha orientadora Rita. Por toda a paciência que teve comigo, pelos ensinamentos nas aulas e conversas. Você me ensinou como fazer ciência (ainda estou aprendendo). Muito obrigado.

Resumo

Utilizando dados de expressão gênica obtidos por RNA-Seq de células únicas de *Mus Musculus*, analisamos o ciclo celular a partir do método do transcriptograma e de análise por componentes principais (PCA). A análise sugere uma classificação das amostras nas diferentes fases do ciclo celular e possibilita propor um ranqueamento pseudo-temporal das amostras. Com dados de grupos de genes reguladores do ciclo, como o complexo ciclina-CDK, validamos biologicamente o ordenamento, uma vez que a sequência temporal proposta pelo ordenamento das amostras dá lugar à evolução esperada da expressão gênica de marcadores de fases do ciclo celular.

Abstract

Using gene expression data obtained from RNA-Seq of single-cell Mus Musculus , we analyzed the cell cycle from the transcriptogram and principal component analysis method (PCA). The analysis suggests a classification of the samples in the different phases of the cell cycle and made possible to propose a pseudo-temporal ordering. With data from groups of cycle-regulating genes, such as the cyclin-CDK complex, we biologically validate the ordering, since the temporal sequence proposed by the ordering of the samples gives rise to the expected evolution of the gene expression of phase markers of the cell cycle.

Sumário

Capítulo 1

Introdução

O ciclo celular é um dos fenômenos mais importantes e desafiadores da biologia, sendo fundamental para o desenvolvimento das células e, portanto, para a existência da vida. Problemas e anomalias resultantes do mal funcionamento do ciclo estão relacionados com a morte das células ou aparição de tumores. A compreensão dos mecanismos de ativação, supressão e regulação do ciclo, associados a seus agentes (organelas, genes, proteínas, ...) são de extrema importância para o desenvolvimento de tratamentos e eventual cura de muitas doenças, como por exemplo o câncer.

O perfil de expressão gênica da célula dá informação sobre quais e em qual quantidade os genes estão sendo expressos. A medida do perfil se dá pela quantificação do RNA mensageiro presente na célula. Este RNA carrega a informação necessária, armazenada nos genes, para que proteínas específicas sejam sintetizadas no ribossomo. Como as proteínas exercem inúmeras funções dentro da célula, ter o conhecimento sobre o RNA transcrito e portanto, sobre o perfil de expressão da célula, possibilita entender quais mecanismos metabólicos estão acontecendo naquele momento. Aliado ao fato que existem muitos genes e que as funções biológicas são executadas por vários produtos gênicos em conjunto, é uma tarefa bastante complicada analisar o perfil de expressão da célula de maneira a elucidar quais funções biológicas estão sendo executadas por um determinado perfil.

Existem muitas técnicas para analisar e obter informações sobre o ciclo celular da célula, através do seu perfil de expressão. Neste trabalho será utilizado uma sequência de técnicas estatísticas diferentes, como por exemplo o transcriptograma e o PCA (análise de componentes principais), para analisar amostras de células únicas (cada amostra corresponde apenas uma célula), na medida que elas passam pelo ciclo celular. Com isso, buscamos aprimorar a análise e obter informações relevantes sobre o sistema, antes desconhecidas.

O método de PCA é útil para identificar as características de um conjunto de amostras que são responsáveis pela variação observada, e então classificar estas características pela quantidade relativa da variância que são responsáveis. No presente trabalho, uma medida consiste na expressão de milhares de genes. Uma característica pode ser tomada como um perfil de expressão destes milhares de genes em quantidades relativas bem determinadas. O PCA auxilia, assim, em identificar perfis de co-expressão gênica cuja variação conjunta descreveu as alterações de expressão gênica observadas. Como mostramos neste trabalho, três desses perfis de co-expressão são responsáveis por mais de 80% da variação observada, desde que seja reduzido o ruído estocástico, inerente às amostras e à técnica experimental. A técnica do transcriptograma contribui para a redução do ruído estocástico, de tal maneira que a razão sinal-ruído fique otimizada.

Após o pré-processamento dos dados pelo transcriptograma, usaremos o método PCA para reduzir a dimensionalidade do sistema (as variáveis correspondentes à genes individuais) para poucas variáveis relevantes ao sistema (alguns perfis de expressão). Essas novas variáveis são as componentes principais. Com estas poucas variáveis (neste trabalho serão utilizadas 3), discriminaremos as amostras pela fase do ciclo celular em regiões, tridimensionalmente.

O objetivo final deste trabalho é obter uma ordem pseudo-cronológica das amostras no ciclo celular. Validamos biologicamente nosso ordenamento, utilizando conhecimentos já existentes sobre o ciclo, como por exemplo, o funcionamento dos complexos ciclina-CDK. Genes esses com papel regulador muito importante dentro da célula.

O trabalho está organizado do seguinte maneira: O capítulo 2 será feito uma revisão sobre o ciclo celular, explicando as principais mecanismos associados a divisão celular, e a divisão em etapas. No capítulo 3 daremos uma explicação básica sobre os métodos de sequenciamento do DNA, especificamente sobre o protocolo RNA-Seq. No capítulo 4 explicaremos os métodos do transcriptograma e da Análise de Componentes Principais utilizados neste trabalho. Por fim, no capítulo 5 apresentaremos a análise, e os resultados obtidos neste trabalho.

Capítulo 2

Revisão de Ciclo Celular

2.1 Ciclo Celular

O ciclo de divisão celular, também conhecido por ciclo celular, é um padrão cíclico que a célula eucarionte realiza com o objetivo de duplicar-se. O ciclo começa com o crescimento da célula, através de um aumento quantitativo das moléculas no seu interior, inclusive de material genético, culminando com a partição do núcleo e do citoplasma em duas células-filhas. A divisão celular é um mecanismo que as células apresentam no desenvolvimento e crescimento de organismos complexos, além de repor células mortas e regenerar partes danificadas dos tecidos e órgãos [1].

O ciclo celular pode ser divido em duas etapas básicas:

- Etapa que a célula cresce e se prepara para a divisão, conhecido como interfase.
- Etapa que ocorre de fato a divisão, do núcleo, conhecida como mitose, seguido da divisão do citoplasma, conhecida como citocinese.

Os processos de crescimento e divisão celular são regulados para que o ciclo transcorra controladamente, assegurando as características essenciais da progênie. Em geral, o tempo de duração do ciclo precisa ser tal que a célula no fim da fase de crescimento tenho o dobro do tamanho original, dando origem a células filhas de mesmo tamanho inicial. Este mecanismo é regulado tanto por produtos gênicos produzidos pela célula como por fatores extracelulares.

Devido à existência de muitos organismos vivos complexos e diferentes, o ciclo celular apresenta diferenças entre os organismos. Entretanto, à medida que aprofundamos o conhecimento sobre o controle do ciclo celular, observamos mais similaridades entre diferentes organismos. Isso mostra uma origem ancestral comum e uma alta conservação evolutiva nos modos de atuação em relação aos genes e proteínas associadas a este processo. Por exemplo, comparando células humanas e de leveduras notamos que transferindo certas proteínas de células humanas para as de leveduras, as proteínas continuam exercendo as mesmas funções anteriores[1].

2.1.1 Interfase - Fase G1

A fase G1 se caracteriza pelo reinício da síntese de RNA de proteínas, que estavam interrompidas durante da fase da mitose (fase M). Nesta fase, a célula começa o seu processo de crescimento, que irá continuar depois

nas fases S e G2. A maior parte das proteínas sintetizadas durante esta fase continuam sendo sintetizadas durante as próximas fases da interfase. Entretanto existem proteínas específicas para esta fase, representando marcadores de G1, como será explicado posteriormente. A maior parte do RNA sintetizado nesta fase é rRNA (RNA ribossômico), chegando por volta de 80% do total sintetizado.

A fase G1 tem interessantes funções preparatórias para a fase subsequente, fazendo a síntese de enzimas imprescindíveis para o funcionamento da fase S, como enzimas catalisadoras da síntese de trifosfatos de desoxirribonucleosídios, enzimas da síntese de DNA polimerase e enzimas dos genes que codificam as histonas. Outra função importante da fase G1, é o papel de decisão celular, podendo a célula continuar o processo do ciclo celular ou então entrar num estado quiescente (G0). A "decisão"pode ser baseada em fatores extracelulares.

2.1.2 Interfase - Fase S

A fase S é marcada pelo início da síntese do DNA e, em geral, entrando nesta fase, não há a possibilidade de interromper o ciclo celular. Nesta etapa, a célula duplica o seu conteúdo de DNA, num processo conhecido como replicação. Na fase G1, a célula apresenta uma quantidade de DNA, enquanto no fim do processo, a célula apresenta o dobro. Esta quantidade permanece até o fim do ciclo celular, de forma a ser repartida igualmente entre as células filhas.

Uma característica importante da replicação do DNA durante a fase S, é o fato de ser o processo semi-conservativo, já que as duas fitas de DNA originais, conhecidas também como parentais, são copiadas originando duas moléculas filhas, que contêm apenas umas das fitas recém sintetizadas. Portanto, cada nova molécula de DNA é uma cópia perfeita de uma molécula preexistente. Além disso, o processo é conhecido como assincrônico, porque a duplicação do DNA não se dá ao mesmo tempo em todas as moléculas de DNA de um núcleo. Regiões específicas do material genético ou genes específicos, começam e terminam em momentos específicos durante o decorrer da fase S.

A duplicação do DNA, em células eucariontes, tem sua origem, simultaneamente, em diferentes pontos do DNA. Cada unidade de replicação é chamado de replicons. As células dos mamíferos apresentam em torno de 20.000 a 30.000 replicons.

Todo este processo de replicação do material genético é realizado por enzimas. Entre as mais importantes, existe a helicase, responsável por quebrar as pontes de hidrogênio do DNA, desenrolando a sua dupla hélice e expondo a cadeia simples do DNA. Desta forma, uma enzima chamada de primase, que é um RNA polimerase especial, tem a função de dar o início do sequenciamento, completando a cadeia de DNA. A partir do início do sequenciamento, a enzima DNA polimerase (DNApol) completa o sequenciamento, sempre com base no molde de DNA. Na medida que a duplicação ocorre, a fibra nucleossômica vai imediatamente se estruturando nas duas novas células DNA, formando novamente o nucleossomo.

A duplicação do DNA é um processo extremamente preciso, existindo uma estimativa de erro menor que 1% nas bases. Isso ocorre devido ao mecanismo conhecido como "proofreading"(teste de leitura), realizado pelo DNAPol, que na medida que adiciona as bases nitrogenadas no filamento de DNA, interrompe o processo se encontrar uma base incorreta, removendo-a.

2.1.3 Interfase - G2

A fase G2 é uma preparação para mitose. Sabe-se na literatura que existem muitos processos ocorrendo, mas não se descobriu todos. Entretanto, sabemos que até a célula entrar na mitose, é necessário que a

replicação do DNA seja totalmente completada e possíveis danos ao DNA tenham sido reparados. Existem controles biológicos na célula, de natureza sensorial, que detectam anormalidades na replicação, e enviam sinais negativos para os sistemas de controle do ciclo, interrompendo a célula de entrar na mitose.

Também nesta fase são sintetizadas proteínas não histônicas (que não estão associadas à decomposição e compactação do DNA), e continua a síntese de proteínas no geral, iniciadas na fase G1, além da síntese de RNA, principalmente os extranucleares.

2.1.4 Mitose

O conteúdo celular já duplicado durante a interfase, é repartido durante a mitose, originando duas células filhas. Ocorre essencialmente dois processos neste período:

- Cariocinese (também conhecido como mitose propriamente dito), na qual ocorre a partilha exata do material nuclear.
- Citocinese, que corresponde à divisão citoplasmática.

A mitose pode ser dividida em 4 etapas: prófase, metáfase, anáfase e telófase. Para este trabalho, não é necessário o conhecimento detalhado destas subfases, já que, os dados que serão analisados das células durante o ciclo celular[2], são catalogados em : G1, S e G2M. As células são classificadas em G2M porque o tempo de duração da mitose é curto, em relação às outras etapas, ficando muito complicado a sua especificação, na forma que foi realizado a identificação.

2.2 Regulação do Ciclo Celular

Certas proteínas e enzimas são responsáveis pelo mecanismos de disparar e coordenar as etapas do ciclo celular, fazendo de suas respectivas concentrações na célula, indicadores da fase do ciclo. Entre elas, os complexos ciclina-CDK serão utilizados neste trabalho.

2.2.1 Ciclina-CDK

A CDK (Cyclin-dependent kinases) é uma família de enzimas quinases, especificamente dependente de ciclinas. As enzimas quinases de proteínas têm a função de fosforizar as proteínas-substratos, consistindo em transferir um grupo fosfato de ATP (Adenosia tri fosfato) para aminoácidos aceptores. Este processo modifica quimicamente as proteínas. Desta forma, diferentes CDKs são ativados ou inativados ao longo do ciclo, gerando padrões cíclicos de fosforilação de proteínas, regulando assim importantes eventos no ciclo.

A atividade das CDKs está relacionada com a associação de proteínas regulatórias, chamadas de ciclinas. O nome ciclina lembra o padrão cíclico de acúmulo e degradação desta proteína ao longo do ciclo celular, com períodos de sintetização na interfase, e sua rápida degradação no fim da mitose. Nas células humanas foram identificadas 10 diferentes ciclinas, denominadas A, B, C, D, e assim por diante. Já as CDKs, identificaram-se mais de 11.

A função de quinase é apenas exercida pela CDK quando está associada à ciclina, formando um dímero (molécula composta por duas unidades similares). Na ausência da ciclina, a CDK é inativa. Portanto, na medida que a ciclina acumula e degrada ao longo do ciclo, a CDK (associada à ciclina) fosforiza proteínas-alvo

específicas, regulando o ciclo celular. A figura 2.1 a seguir mostra a concentração das ciclinas A, B, D e E ao longo do ciclo.

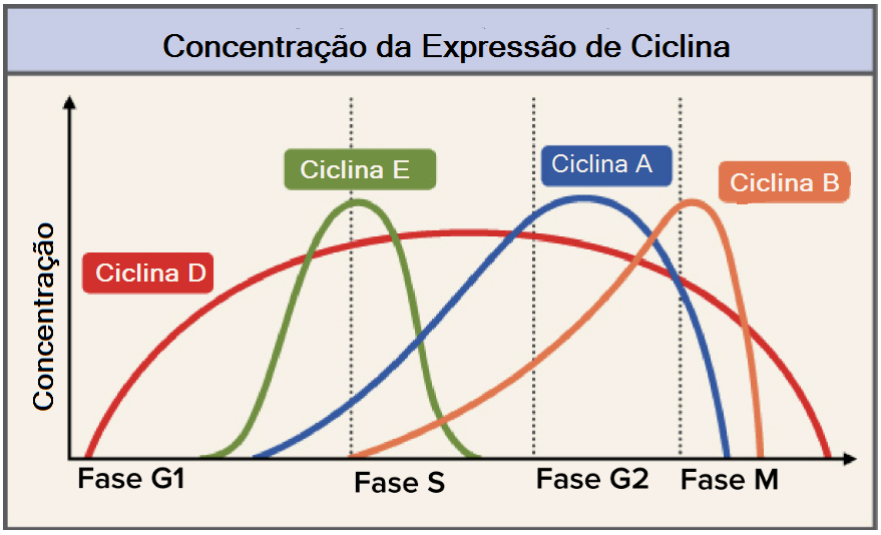

Figura 2.1: Concentração das Ciclinas ao longo do ciclo celular[3].

As ciclinas A, B, D e E são muito importantes no ciclo celular, coordenando funções e representando marcadores da atividade metabólica celular:

- Ciclina D: Está presente durante todo o ciclo celular, em maior ou menor escala. Tem sua expressão iniciada na fase G1 e término na mitose. Apresenta ligada às CDKs 4 e 6 no ciclo celular.
- Ciclina E: Apresenta um pico de concentração na transição da fase G1/S, determinando o início da duplicação do DNA. Apresenta-se ligada à CDK 2 no ciclo celular.
- Ciclina A: Inicia sua expressão no fim da fase G1, aumentando até a fase G2, local que apresenta um pico. Apresenta muitas funções ao longo do ciclo, dependendo da sua ligação com determinada CDK.
- Ciclina B: Tem início da expressão no fim da fase G1, sendo muito importante na mitose. Apresenta uma grande concentração na transição G2/M, com seu pico na mitose.

Neste trabalho analisamos dados de RNA-Seq que dão informação sobre o perfil dos RNA mensageiros presentes na célula. Assim, células em diferentes estágios do ciclo apresentam diferentes perfis de mRNA e, por meio destas diferenças, podemos classificar e obter informação sobre a dinâmica deste processo.

Capítulo 3

Sequenciamento do DNA - RNA-Seq

3.1 Revisão Histórica

As técnicas de sequenciamento do DNA iniciaram na década de 70, com trabalhos desenvolvidos por Sanger e seus colaboradores[4] e por Maxam e Gilbert[5]. Como a técnica desenvolvida por Sanger necessitava de um manuseio menor de produtos químicos tóxicos e a utilização de radioscópios em relação à desenvolvida por Maxam e Gilbert, o sequenciamento Sanger (como ficou conhecido o método), se tornou a principal técnica de sequenciamento do DNA pelos próximos 30 anos.

O Sequenciamento Sanger, de forma simplificado, é um método que consiste em adicionar nucleotídeos modificados - didesoxiribonucleotídeos (ddNTP's)- durante o processo de replicação do DNA. Estes nucleotídeos apresentam uma modificação em relação aos nucleotídeos naturais, que é o fato de não possuírem o grupo OH livre no carbono 3' da pentose. Esta alteração faz com que o processo de replicação do DNA pare quando estes nucleotídeos modificados se ligam à fita de DNA, já que o próximo nucleotídeos não têm como se ligar na sequência, devido à ausência do grupo OH livre. Desta forma, a adição do nucleotídeo modificado em momentos diferentes da replicação do DNA, obteremos diferentes resíduos na fita de DNA em relação ao mesmo DNA analisado. A partir do continuo avanço tecnológico, aliado a um investimento em automação e paralelização dos processos, foi possível melhorar o rendimento e produtividade nas técnicas de Sanger, possibilitando um sequenciamento completo do genoma humano em 2003[6].

O Instituto Nacional de Pesquisa do Genoma Humano (NHGRI) dos EUA, lançou uma iniciativa de financiamento de projetos referentes ao sequenciamento do genoma humano em 2004. O principal objetivo do projeto era reduzir para U$1000,00 o custo do sequenciamento do genoma humano em 10 anos.

Este grande investimento na pesquisa sobre o genoma humano, fez surgir uma nova geração de técnicas de sequenciamento do genoma (NGS), entre elas, as principais são: Pirosequenciamento com detecção de pirofosfato (454 – Roche), Sequenciamento por ligação (SOLiD), Metodologia de semicondutores (Ion), Sequenciamento por síntese (Illumina), Sequenciamento de moléculas únicas (Pacific Biosciences e o Oxford Nanopore), entre outras. Estas novas técnicas permitiram um sequenciamento direto e paralelo de bilhões de moléculas de DNA, aliado à redução substancial de amostras necessárias, podendo chegar a amostras de células únicas (single-cell), além do objetivo principal, uma grande redução do custo por nucleotídeo sequenciado.

Com este grande avanço das técnicas de sequenciamento do genoma (não só humano, mas de diferentes espécies), surge um novo protocolo de sequenciamento em 2008, chamado de RNAseq (RNA sequencing)[7]. Neste caso, o resultado da análise é o perfil de RNA mensageiros presentes nas células (não mais o sequenciamento do DNA). Este protocolo foi utilizado para obter os dados da expressão gênica de 288 amostras de células únicas de *Mus Musculus*[2] que serão utilizados neste trabalho.

3.2 RNASEQ

O método de RNAseq é uma abordagem que utiliza tecnologias de sequenciamento de DNA para traçar um perfil do transcriptoma da célula. O transcriptoma é o resultado da medida de expressão gênica que resulta no conjunto dos transcritos de determinada célula, em um instante específico de desenvolvimento e condições biológicas. As informações obtidas pelo transcriptoma são importantes para entender o metabolismo celular. Com este tipo de abordagem é possível catalogar todas as espécies de transcritos como por exemplo mRNAs (RNA mensageiro), RNA não codificantes (RNA que não é traduzida em proteínas, como por exemplo RNA de transferência e RNA ribossomial), micro e pequenos RNAs. Assim , o resultado de uma medida de RNASeq possibilita determinar a estrutura transcricional dos genes e quantificar os níveis de expressão de cada gene em diferentes condições[7].

Esta nova abordagem apresenta vantagens em relação a antigos métodos utilizados para obter o sequenciamento e o transcriptoma da célula. Entre estes métodos, existe o método baseado em hibridização e sequenciamento Sanger por cDNA. No método baseado em hibridização é feito uma análise DNA-DNA medindo o grau de semelhança genética entre genomas completos. Este grau de semelhança pode ser entre indivíduos da mesma espécie ou de espécies diferentes. A partir de uma amostra base, com os fragmentos de DNAs de fita simples já identificados e marcados (de forma radioativa), verificam-se quais sequências de DNA de uma amostra de análise são ligados aos fragmentos de DNA marcados. Contando a quantidade de sequências de DNA da amostra de análise que estão ligadas a determinadas sequências de DNA da amostra de base, obtêm-se quais genes estão sendo mais ou menos expressos naquela amostra.

O método de RNAseq apresenta vantagens em relação ao método de hibridização e sequenciamento Sanger. A resolução dos genomas no método baseado em hibridização gira em torno de 100 bp (pares de base), enquanto no RNAseq a resolução é um simples par. Outro problema, é a ineficácia de distinguir diferentes isoformas e expressão alélica. De forma prática, o sequenciamento Sanger necessita de uma quantidade grande de RNA e um custo alto para mapear transcriptomas de grandes genomas, em relação ao método de RNAseq.

O sequenciamento Sanger também é ineficaz em distinguir diferentes isoformas e expressão alélica, além de não conseguir mapear de forma simultânea regiões transcritas e expressão gênica.Em razões práticas para o experimento, é necessário uma utilização grande de RNA e um custo alto em mapear o transcritomas de grandes genomas em relação ao RNAseq.

3.3 PREPARAÇÃO DA AMOSTRA E CONTAGEM DO CDNA

O experimento utilizando o protocolo de sequenciamento por RNA (RNAseq) começa isolando o RNA de uma amostra que se procura analisar. Este RNA isolado precisa ter uma qualidade suficiente para conseguir produzir uma biblioteca de sequenciamento. A produção de uma biblioteca de sequenciamento de baixa qualidade pode gerar conclusões biológicas errôneas. Uma medida para avaliar a qualidade do RNA isolado é o

RNA Integrity Number (Número de integridade do RNA) também conhecido como RIN[8]. A escala do RIN vai de 0 a 10, considerando o RNA com um número menor que 6 de baixa qualidade. A medida RIN consiste em, sabendo que os ácidos nucléicos apresentam carga negativo, fazer uma análise eletroforese do RNA.

O segundo passo do protocolo de sequenciamento por RNAseq consiste em elaborar uma biblioteca de RNA. Entretanto, antes da construção da biblioteca, é preciso fazer algumas preparações nas amostras de RNA que se quer analisar. Dentro da célula, a maior parte do RNA é de origem ribossômica (rRNA), chegando em média a 95%. Quando se procura obter a expressão gênica ou sequenciar o DNA, o rRNA pode consumir muito a leitura do RNA total, limitando a cobertura da sequência e, portanto, limitando a detecção de RNA menos frequentes. Um dos processos mais eficientes para resolver o problema, é enriquecer a amostra com RNA mensageiro (mRNA). O procedimento adequado para eliminar a contribuição do rRNA depende da finalidade do experimento.

Esta biblioteca de RNA pode variar de acordo com a espécie da amostra e também com o tipo de sequenciamento utilizado. Para criar esta biblioteca, precisamos isolar o RNA que se deseja analisar, e por um processo de transcriptase reversa, transformar este RNA em cDNA. Este cDNA será fragmentado e multiplicado de forma aleatória, e ligados a adaptadores de sequenciamento. É interessante salientar que esta etapa básica de preparação da biblioteca de RNA pode ter variações dependendo do RNA que ser analisar, de acordo com os objetivos do experimento.

No experimento de RNAseq é muito importante selecionar o material biológico adequado para fazer a biblioteca de RNA. Esta escolha sempre dependerá do objetivo do experimento, já que cada tipo de célula ou tecido pode apresentar funções específicas e diferentes. Outra consideração importante é a escolha do momento temporal para realizar a preparação da biblioteca, devido ao fato de que, dependendo do estágio de desenvolvimento da célula, o valor da expressão gênica será diferente ao longo do tempo. Esta última consideração é muito importante em relação ao experimento analisado neste trabalho.

Sabendo que os cDNAs da amostra, após todos os passos anteriores, são informações a respeito do mRNA presente na célula: contando este cDNA obtém-se a dinâmica da expressão gênica na amostra. Isso acontece porque o mRNA que vem do núcleo da célula está levando para o citiplasma a informação de quais proteínas e processos serão realizados. Assim, contando as cDNAs da amostra conseguimos obter uma "fotografia"dos processos que estão ocorrendo.

Por fim, estas moléculas são sequenciadas e contadas. Existem muitas tecnologias de sequenciamento de alto rendimento utilizado neste processo, como Illumina IG[9], Applied Biosystems SOLiD[10] e Roche 454 Life Scienc[11]. Após sequenciar, as leituras resultantes são alinhadas a um genoma de referência ou transcripto de referência. Estes genes alinhados reproduzem um mapa de transcrição, que dá informação sobre o nível de expressão de cada gene, de acordo com a quantidade de cada gene alinhado.

Capítulo 4

Metodologia

4.1 Transcriptograma

O método do transcriptograma é uma técnica de análise estatística de dados[12] desenvolvida pelo grupo do Laboratório de Estruturas da Celulares do Instituto de Física da UFRGS. O método consiste em apresentar e analisar dados de expressão de genoma inteiro, onde são suavizados os ruídos estocástico e biológico inerentes, respectivamente, à técnica de medida e à variação típica de sistemas biológicos. Esse tratamento estatístico dos dados de expressão possibilita o aumento da razão sinal-ruído das medidas.

O método de transcriptograma baseia-se no ordenamento em uma lista dos genes de um organismo de tal maneira a agrupar genes cujos produtos participam de uma mesma função biológica. Este ordenamento é não supervisionado e lança mão da informação de associações proteicas disponibilizada pelo STRING, um consórcio de universidades e centros europeus, incluindo o European Molecular Biology Laboratory (https://string-db.org/). A projeção das medidas de expressão sobre esta lista ordenada possibilita que uma média de janelas de tamanho pré-definidos otimizem a razão sinal-ruído, melhorando a reprodutibilidade e aumentando a sensibilidade das medidas[13].

Embora na literatura científica já existam métodos de análise de expressão gênica que obtêm a média de expressão tomada sobre um conjunto de genes funcionalmente relacionados, o diferencial do método do Transcriptograma está na maneira de definir os conjunto de genes.

Nas próximas seções explicaremos o método de transcriptograma e a sua aplicação na análise dos dados de expressão gênica das células únicas de *Mus Musculus*, passando pelo ciclo celular[2].

4.1.1 Lista de Genes Ordenados - Método da Função Custo

Antes de realizar a média da expressão gênica sobre um conjunto de dados, em uma determinada vizinhança (raio), é necessário ordenar uma lista de genes de uma forma adequada, de maneira que a probabilidade de que genes funcionalmente relacionados decaia exponencialmente com a distância na lista. O método utilizado para ordenar esta lista de forma adequada foi o Método da Função Custo (CFM).

Os dados sobre associação proteína-proteína, utilizados para a elaboração da lista ordenada, está disponível publicamente no banco de dados STRING[14]. O banco de dados fornece uma lista de pares de proteínas que são classificadas como associadas segundo sete métodos de inferência, que englobam, por exemplo, interação

física, participação em uma mesma rota metabólica ou ´text mining´. Para cada associação proteica nesta lista o banco apresenta um valor de confiança que indica a confiança que estas duas proteínas façam parte de uma mesma rota metabólica. O valor de confiança vai de 0 até 1 para cada par de produto gênico. Neste trabalho, será considerado que 2 pares de genes estejam correlacionados se o valor de confiança for maior que $0,8$.

Com os dados dos produtos gênicos associados, montamos uma matriz M de forma que:

- Se a confiança da associação for maior que $0,8$ então $M_{i,j} = 1$.
- Se a confiança da associação for menor que $0,8$ então $M_{i,j} = 0$.

A figura a seguir, apresenta esta matriz inicial antes de realizar o ordenamento:

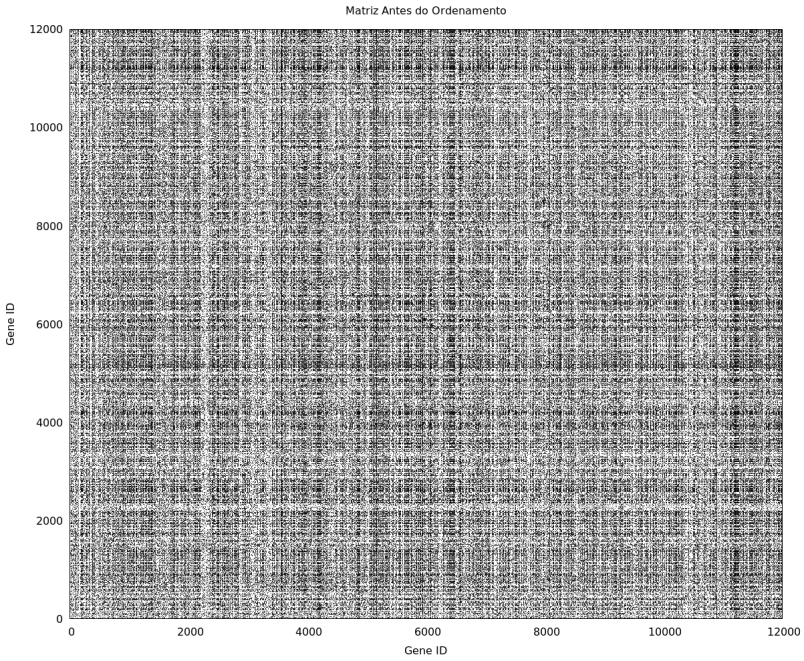

Figura 4.1: Matriz A antes do ordenamento

Como pode ser visto pela figura 5.1, esta matriz representa a associação de pares de genes que, antes do ordenamento, estão aleatoriamente distribuídos. Na figura, os pontos em preto referem-se aos pares de genes que estão associados e os pontos em branco, aos pares de genes que não estão associados.

O primeiro passo para ordenar os genes, é calcular o custo do estado inicial da matriz, usando a seguinte equação:

$$H = \sum_{i=1}^{N} \sum_{j=1}^{N} \left|i-j\right|^{\alpha} \left(\left|M_{i,j} - M_{i+1,j}\right| + \left|M_{i,j} - M_{i-1,j}\right| + \left|M_{i,j} - M_{i,j+1}\right| + \left|M_{i,j} - M_{i,j-1}\right|\right) \tag{4.1.1}$$

Esta função custo apresenta dois termos:

- $\left|i-j\right|^{\alpha}$, que depende da distância dos genes na posição i e j na lista. Nota-se que se $\alpha > 0$ e $M_{i,j} = 1$ e portanto os genes nas posições i e j estão associados, a função custo H é menor se i e j ficarem mais próximos.

- O termo entre parênteses é maior quando os elementos dos vizinhos de $M_{i,j}$ não são associados. Quando os vizinhos de $M_{i,j}$ são associados, o valor entre parênteses diminui.

Analisando a equação (3.1.1), α controla a força do primeiro termo. Na literatura científica, para analisar o ciclo celular, já foi utilizado o valor de $\alpha = 1$[15]. Este valor será utilizado neste trabalho.

Ordenar a lista de forma que os genes que estão associados fiquem próximos é equivalente a diminuir o valor da função custo H. Para diminuir o valor de H é realizado uma simulação de Monte Carlo [16]. Em cada passo da simulação, troca-se as posições de dois genes, de forma aleatória. Em cada troca, é novamente calculado o valor da função custo H, e feito a diferença entre o valor custo final em relação ao passo anterior:ΔH. Calculado a diferença, é realizado o seguinte critério:

- Se $\Delta H \leq 0$, então esta troca de posição de genes é mantida.
- Se $\Delta H > 0$ a troca é aceita com uma probabilidade exp $\left(-\frac{\Delta H}{T}\right)$ de ocorrer.

O parâmetro T no método de Monte Carlo, é semelhante à temperatura. O valor de T inicial corresponde a $0,01\%$ do valor inicial da função custo H e a cada 100 passos, este valor é reduzida à metade. A possibilidade de ainda ocorrer a troca mesmo quando aumenta o valor da função custo, é importante para evitar o estados metaestáveis, procurando sempre um mínimo global. Este processo é conhecido como *Simulated Annealing*.

Depois de realizados um número de suficiente de passos (utilizamos 40 mil), obtemos o seguinte ordenamento, observados pela matriz final A obtida:

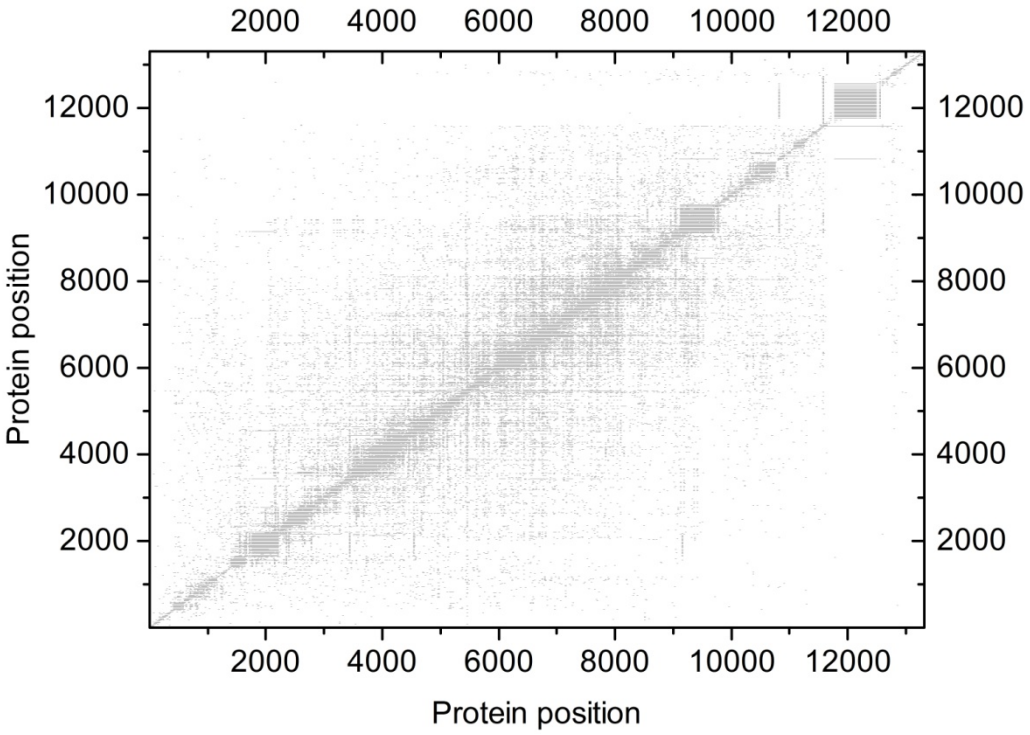

Figura 4.2: Matriz A depois do ordenamento

Analisando a figura 4.2 notamos que, depois do ordenamento, as proteínas que estão associadas estão próximas, já que os pontos da matriz que indicam associação proteína-proteína localizam-se em torno da diagonal.

4.1.2 Médias e Transcriptograma

Tomando a lista ordenada dos genes, o transcriptograma de uma medida de expressão gênica de genoma inteiro é obtido da seguinte forma:

1. atribui-se ao gene localizado na posição i da lista ordenada o valor da expressão gênica t_i;
2. realiza-se uma média sobre a expressão de cada gene e as expressões de seus r vizinhos à direita e seus r vizinhos à esquerda sobre a lista ordenada, onde r é um parâmetro previamente escolhido;
3. o valor da média é atribuído ao gene na posição i como sendo o seu valor de transcriptograma τ_i.

O valor do transcriptograma τ_i para o gene na posição i é, então, dado pela seguinte equação:

$$\tau_i = \langle t_i \rangle_{w_i} = \frac{\sum_j H\left(r - d_{i,j}\right) t_i}{\sum_j H\left(r - d_{i,j}\right)} \quad , \tag{4.1.2}$$

onde $w_i = 2r + 1$ é a janela ou região do raio r sobre a qual as médias são realizadas. Por fim, $d_{i,j}$ corresponde à distância do gene na posição i em relação ao gene na posição j e H é dado por

$$H(x) = \begin{cases} 0, & \text{se } x < 0 \\ 1, & \text{se } x \geq 0 \end{cases} .$$

Existem muitos critérios possíveis para a escolha do raio do transcriptograma. Neste trabalho, o critério utilizado foi a qualidade da classificação pelo método de análise principais (PCA) quanto à fase no ciclo das células cujas medidas de RNA-Seq foram realizadas. Como explicado posteriormente, o raio que utilizamos foi o de $r = 30$.

4.1.3 Lógica Biológica

O agrupamento dos genes por sua função biológica pode ser verificado pela projeção da lista dos genes que estão associados a diferentes termos do Gene Ontology: Biological function (GO:BP)[17] ou como uma rota metabólica do KEGG[18]. Esta projeção é feita da seguinte maneira: para cada termo da GO:BP ou rota metabólica do KEGG, atribui-se o valor 1 ou 0 para cada posição da lista ordenada dependendo se o gene naquela posição está ou não associado ao termo ou rota. Em seguida, para cada posição da lista ordenada, calula-se a média sobre um raio pré definido (neste caso, 30). Como resultado obtemos perfis que apresentam um máximo na região da lista onde os genes/proteínas da função biológica estão concentrados.

A Figura 4.3 apresenta a perfil de funções biológicas escolhidas, evidenciando o agrupamento das mesmas. Mais ainda, o ordenamento obtido como explicado acima agregou genes/proteínas primeiramente envolvidos com metabolismo de energia (à esquerda, em tons de verde), então genes envolvidos com RNA e tradução, seguido por termos que estão envolvidos com processamento de DNA e ciclo celular. Mais para a parte central, há os genes que participam em rotas de diferenciação celular, então as envolvidas com citoesqueleto e interação com o meio extra celular. Finalmente, metabolismo de drogas enquanto o grande pico em cinza, à extrema

direita da lista, corresponde a receptores olfactórios. Esse agrupamento de genes/proteínas envolvidos em uma mesma função biológica faz com que médias de janelas de dados de expressão tenham um sinal correlacionado, enquanto o ruído permanece descorrelacionado: médias sobre intervalos na lista, portanto, otimizam a razão sinal-ruído.

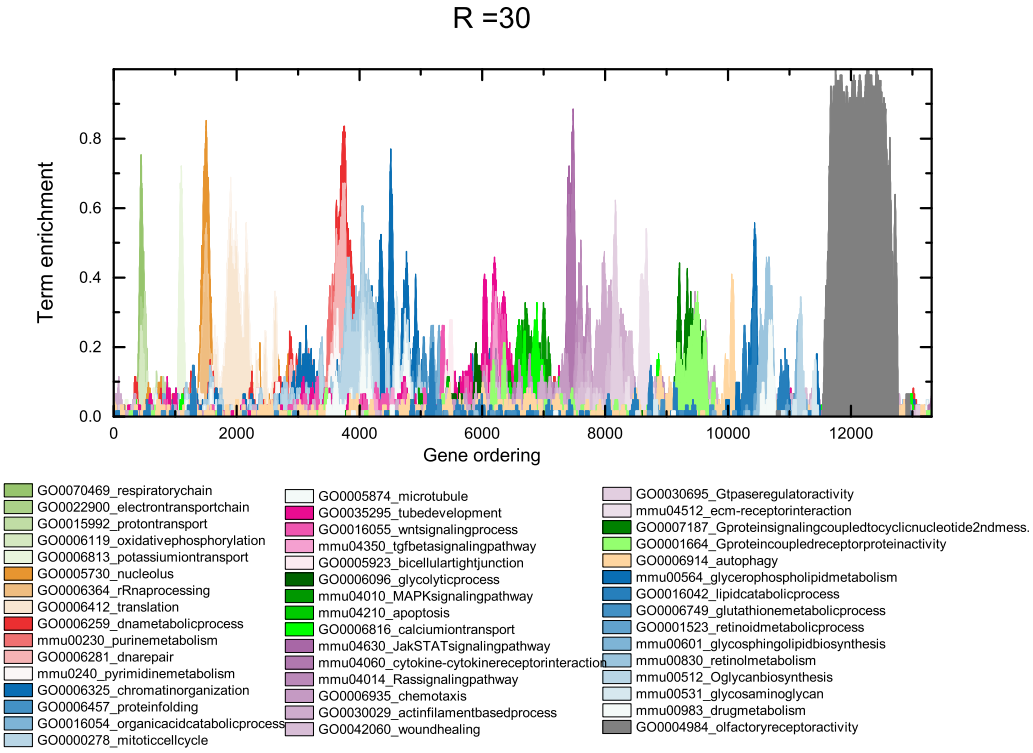

Figura 4.3: Distribuição das funções biológicas ao longo da lista ordenada

4.2 Análise de Componentes Principais - PCA

A ideia central do método de análise por componentes principais (PCA)[19] é reduzir a dimensionalidade da descrição de um sistema, a partir de um conjunto de dados que consistem de um grande número de variáveis para cada amostra. Este objetivo é alcançado pela rotação e translação no espaço das muitas variáveis que inicialmente descrevem as amostras, de tal maneira que os novos eixos são escolhidos pela variação que contêm. Para que a PCA seja efetiva, a informação contida nas medidas concentra-se em um número reduzido de componentes onde a variação do sistema é significativa. As novas componentes são ordenadas de forma decrescente na quantidade de variação que apresentam. Estas novas variáveis são chamadas de componentes principais.

Nas próximas seções será apresentado como obter as componentes principais de forma algébrica.

4.2.1 Álgebra

Para encontrar as componentes principais de forma algébrica[20], vamos supor um experimento com muitas amostras e que a caracterização de cada amostra é feita pela obtenção de várias medidas (*features*). Exemplo é o transcriptoma obtido pela medida de RNASeq de diferentes amostras. Podemos representar tal conjunto de dados por uma matriz $\mathbf{X}$, com n colunas, que representam as amostras, e m linhas, que representa as variáveis (expressão dos genes). Para transformar a matriz $\mathbf{X}$, $m \times n$ em uma outra matriz $\mathbf{Y}$, também m x n, temos que obter a matriz $\mathbf{P}$, tal que:

$$\mathbf{Y} = \mathbf{PX} \tag{4.2.1}$$

A equação 4.2.1 é simplesmente uma mudança de base.

Organizando as linhas de $\mathbf{P}$ de forma que $\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{p_3}$,.., $\mathbf{p_m}$ sejam vetores linha de $\mathbf{P}$, então a matriz $\mathbf{P}$ pode ser escrita como

$$\mathbf{P} = \begin{pmatrix} \mathbf{p_1} \\ \mathbf{p_2} \\ \vdots \\ \mathbf{p_m} \end{pmatrix},$$

enquanto a matriz $\mathbf{X}$ pode ser escrita usando os vetores colunas $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$,.., $\mathbf{x_m}$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x_1} & \mathbf{x_2} & \cdots & \mathbf{x_n} \end{pmatrix}$$

Podemos agora interpretar a equação 4.2.1 da seguinte forma:

$$\mathbf{PX} = (\mathbf{Px_1}\, \mathbf{Px_2}\, \mathbf{Px_3} \ldots \mathbf{Px_n}) = \begin{pmatrix} \mathbf{p_1x_1} & \mathbf{p_1x_2} & \cdots & \mathbf{p_1x_n} \\ \mathbf{p_2x_1} & \mathbf{p_2x_2} & \cdots & \mathbf{p_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{p_mx_1} & \mathbf{p_mx_2} & \cdots & \mathbf{p_mx_n} \end{pmatrix} = \mathbf{Y}$$

É interessante notar que $p_i x_j \in \mathbb{R}^m$ e, assim, esta operação é apenas o produto interno entre $\mathbf{p_i x_j}$. Com isso, se observa que a matriz original $\mathbf{X}$ está sendo projetada nas colunas de $\mathbf{P}$, e as linhas de $\mathbf{P}$, $\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{p_3}$,.., $\mathbf{p_m}$ representam uma nova base para representar $\mathbf{X}$. Fazendo algumas considerações importantes e um desenvolvimento adequado, verifica-se que as linhas de $\mathbf{P}$ são as componentes principais de $\mathbf{X}$.

No método de PCA, descorrelaciona-se as variáveis, encontrando uma base de representação na qual as novas direções são aquelas que a variância é concentrada nas primeiras componentes.

Considere agora um vetor qualquer no espaço das variáveis que definem o experimento analisado, dado por $\tilde{\mathbf{r}} = (\tilde{\mathbf{r_1}}, \tilde{\mathbf{r_2}}, \tilde{\mathbf{r_3}}, \ldots, \tilde{\mathbf{r_n}}) \in \mathbb{R}^n$. Considere também que $\mu_{\tilde{r}}$ é o valor médio das componentes de $\tilde{\mathbf{r}}$. Subtraindo este valor médio de cada uma das componentes, teremos o vetor $\mathbf{r} = (\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}, \ldots, \mathbf{r_n}) \in \mathbb{R}^n$, que apresenta média $\mu_r = 0$. Agora, podemos calcular a variância σ_r^2 das componentes de $\mathbf{r}$ pelo valor de seu módulo quadrado, isto é,

$$\sigma_r^2 = \frac{1}{n}\mathbf{r}\mathbf{r}^T \ . \tag{4.2.2}$$

Definindo um novo vetor $\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \mathbf{s_3}, \ldots, \mathbf{s_n}) \in \mathbb{R}^n$, que apresenta também $\mu_s = 0$, podemos generalizar a

ideia de variância, e definir a covariância de **r** e **s**, expresso da seguinte forma:

$$\sigma_{rs}^2 = \frac{1}{n-1}\mathbf{r}\mathbf{s}^T . \tag{4.2.3}$$

Desta forma, a variância pode ser interpretada como um caso particular da covariância. A covariância informa qual é o grau de interdependência ou inter-relação entre duas variáveis aleatórias. Agora, calculando a covariância da variável **X**:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{pmatrix} \in \mathbb{R}^{m\times n} . \tag{4.2.4}$$

Analisando a equação anterior, a matriz **X** pode ser vista como m vetores linha, com n componentes cada, lembrando que nesta representação, m corresponde ao número de variáveis e n, ao número de amostras. Portanto, o vetor $\mathbf{x_i}$ corresponde ao vetor das n amostras, para a i-ésima variável. Sabendo que $\mathbf{x_i}^T \in \mathbb{R}^n$, calculando o produto pela equação 4.2.3 temos que:

$$C_X = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \frac{1}{n-1}\begin{pmatrix} \mathbf{x_1}\mathbf{x_1^T} & \mathbf{x_1}\mathbf{x_2^T} & \cdots & \mathbf{x_1}\mathbf{x_m^T} \\ \mathbf{x_2}\mathbf{x_1^T} & \mathbf{x_2}\mathbf{x_2^T} & \cdots & \mathbf{x_2}\mathbf{x_m^T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m}\mathbf{x_1^T} & \mathbf{x_m}\mathbf{x_2^T} & \cdots & \mathbf{x_m}\mathbf{x_m^T} = \end{pmatrix} \in \mathbb{R}^{m\times m} ;. \tag{4.2.5}$$

Analisando os termos de $\mathbf{C_X}$, notamos que existem todos os possíveis pares de covariância entre os vetores x_i, sendo que a diagonal principal representam a variância de x_i, e os termos fora da diagonal representa a covariância da i-ésima variável entre diferentes amostras. A matriz C_X é chamada de matriz de covariância.

Agora, precisamo, a partir da transformação linear, dada pela equação 4.2.1, obter a matriz **Y** que se relaciona com a matriz de covariância $\mathbf{C_X}$. Como já foi dito anteriormente, a covariância pode ser considerada uma medida da correlação entre duas variáveis. Assim, partindo do pressuposto inicial, que o método de PCA busca que as variáveis da matriz transformada sejam mais descorrelacionados possível, as covariâncias de diferentes variáveis na matriz $\mathbf{C_Y}$ precisam ser o mais próximo de zero. Entretanto, quanto maior a variância, mais informação podemos obter sobre a dinâmica do sistema. Logo, a construção da matriz de variância segue as seguintes propriedades:

- Maximizar o sinal, medido pela variância. Isso equivale a maximizar os valores na diagonal principal.
- Minimizar a covariância entre variáveis. Isso equivale a minimizar os valores fora da diagonal principal.

Analisando as características da matriz de covariância, percebemos, que o objetivo é encontrar uma matriz de transformação **P**, tal que a matriz de covariância $\mathbf{C_Y}$ seja diagonal. Supondo que a matriz **P** é ortonormal, podemos utilizar propriedades de álgebra linear para encontrar a solução. Utilizando a equação 4.2.1 e 4.2.5, temos o seguinte:

$$\mathbf{C_Y} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1}(\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T = \frac{1}{n-1}(\mathbf{P}\mathbf{X})\left(\mathbf{X}^T\mathbf{P}^T\right) = \mathbf{P}\left(\mathbf{X}\mathbf{X}^T\right)\mathbf{P}^T$$

Definindo $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, temos:

$$\mathbf{C_Y} = \frac{1}{n-1}\mathbf{P}\mathbf{S}\mathbf{P}^T \tag{4.2.6}$$

Como $\left(\mathbf{X}\mathbf{X}^T\right)^T = \left(\mathbf{X}^T\right)^T(\mathbf{X})^T = \mathbf{X}\mathbf{X}^T$, então $\mathbf{S}$ é uma matriz simétrica $m \times m$. Pelas propriedades de álgebra linear, temos que toda matriz simétrica é ortogonalmente diagnolizável, e portanto:

$$\mathbf{S} = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{4.2.7}$$

sendo que $\mathbf{E}$ uma matriz ortonormal $m \times m$, cujas colunas são autovetores ortonormais $\mathbf{S}$, e $\mathbf{D}$ uma matriz diagonal que possui os autovalores de $\mathbf{S}$. Escolhendo de forma adequada, $\mathbf{P} = \mathbf{E}^T$, utilizando as equações 4.2.1 e 4.2.6, obtém-se:

$$\begin{aligned}\mathbf{C_Y} &= \frac{1}{n-1}\mathbf{P}\mathbf{S}\mathbf{P}^T \\ &= \frac{1}{n-1}\mathbf{E}^T\left(\mathbf{E}\mathbf{D}\mathbf{E}^T\right)\mathbf{E} = \frac{1}{n-1}\mathbf{D},\end{aligned}$$

dado que $\mathbf{E}\mathbf{E}^T = \mathbf{I}$ corresponde a uma matriz identidade $m \times m$. Portanto com os autovalores da matriz $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ podemos formar a matriz $\mathbf{D}$, que é diagonal. Os autovetores associados a $\mathbf{D}$ formam a matriz E. Os valores na diagonal de $\mathbf{D}$ seguem em ordem decrescente dos autovalores. As direções dos autovetores, que apresentam os maiores autovalores, correspondem às componentes principais deste conjunto de dados. Notando que estas direções estão associados a maiores variâncias, de forma que estas componentes contêm informações importantes sobre o conjunto de dados.

Logo, para encontrar as direções, neste conjunto de dados, que apresentam a maior variância, e portando têm papel importante para entender a dinâmica do sistema, precisamos resolver um problema de autovalores e autovetores da matriz de covariância.

Neste trabalho utilizamos a linguagem R para obter a análise de componentes principais do conjunto de dados que analisamos.

Capítulo 5

Análise

5.1 As amostras e sua análise original

Neste trabalho utilizaremos dados da medida de expressão gênica de células T de *Mus Musculus*. Estas medidas estão disponibilizadas no repositório ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) sob o código E-MTAB-2805. A análise destes dados foi publicada em [Buettner, et al, 2015][2]. Os arquivos utilizados apresentam a contagem de *reads* por gene, assim como o comprimento dos genes alinhados, a contagem dos genes não alinhados, ambíguos, de baixa qualidade e alinhados com genes não identificados.

Buettner e colaboradores tinham o objetivo de propor uma abordagem computacional, chamado de Modelo de Variável Latente de Células Únicas (scLVM), que facilita a identificação de subpopulações celulares. Para tal, foram utilizados dados de células em diferentes fases do ciclo celular (96 células da fase G1, S e G2M). Durante o ciclo celular, ocorrem grandes mudanças metabólicas, que modifica os perfis de expressão gênica, de tal forma que o ciclo pode mascarar outras diferenças fisiologicamente importantes.

A publicação[2] inicialmente identificou os genes cujas expressões variam com o estágio do ciclo celular, tanto marcadores gênicos já identificados na literatura, como aqueles cujas expressões apresentaram uma correlação significativa com estes marcadores do ciclo celular. Foram identificados 2881 genes (44% do total) que estão correlacionados com genes do ciclo celular. Isso significa que há muitos genes que variam a intensidade de expressão ao longo do ciclo celular, modificando o perfil de expressão obtido. Depois de encontrados os genes correlacionados com o ciclo celular, os autores descontam esta variação devido ao ciclo. Assim, ao comparar o transcriptoma de células de classes diferentes, a variação não desejada devido ao ciclo interfere menos nas análises estatísticas para a determinação de genes ou conjuntos de genes diferencialmente expressos nas classes.

Uma das formas de validação da publicação, é de fazer a análise por componentes principais (PCA) antes e depois de descontar a contribuição do ciclo na expressão. Foi mostrado que o scLVM consegue remover de forma significativa a contribuição do ciclo celular. Também foi feita uma análise para validar o método, utilizando genes relacionados aos linfócitos TH2. Mostrou-se que, antes de realizar a correção referente ao ciclo celular, é impossível obter subpopulações de genes relacionados ao TH2.

Neste trabalho, diferentemente da publicação [2] que desconsiderou os efeitos, estudaremos aprofundadamente o ciclo celular. Daremos um passo na discriminação das células referentes ao momento no ciclo celular, propondo um ranqueamento pseudo-cronológico destas amostras.

5.2 A ANÁLISE POR TRANSCRIPTOGRAMAS E PCA

5.2.1 Controle de Qualidade

Antes da realização da análise dos dados por transcriptograma e PCA, é necessário passar as 288 amostras (96 amostras para cada fase no ciclo celular) por um controle de qualidade. Existem diferentes formas de controle de qualidade. A que será utilizada neste trabalho corresponde à realizada por Buettner et. al[2]. O controle de qualidade precisa avaliar duas situações:

1. Amostras de baixa qualidade, típicos de experimento por RNASeq.
2. Classificação não confiável das amostras na fase do ciclo celular.

Para avaliar a qualidade do experimento de RNASeq, foram utilizados os seguintes critérios :

1. Contagem Total dos Genes na Amostra > 5 milhões
2. Contagem dos genes alinhados/Contagem total > 0.2
3. Número de Genes com contagens > 6 mil
4. Contagem dos genes ERCC/Contagem dos genes Alinhados < 0.6
5. Contagem dos genes endógenos/Contagem dos genes Alinhados = 1 - Contagem dos Genes ERCC/Contagem dos genes Alinhados
6. Contagem dos genes Mitocondrias/Contagem dos Genes Endógenos < 0.15

Os ERCC(External RNA Controls Consortium) é um método que consiste em colocar trancritos controles na amostra, de outros organismos, antes da realização do RNASeq. Neste caso, a multiplicação das leituras pode ser controlada pela quantidade final destes transcritos controles.

Após aplicar o primeiro controle de qualidade, restaram 81 células na fase G1, 76 células na fase S e 89 células na fase G2M. O segundo controle de qualidade, para verificar a confiabilidade na classificação das amostras no ciclo celular, consiste em, para cada fase, excluir as amostras cuja diferença entre a fração da contagem dos genes endógenos com relação à contagem dos genes alinhados exceder um desvio absoluto da mediana (MAD).

O desvio absoluto da mediana (MAD) é uma medida da dispersão estatística.. Para os dados distribuídos de forma normal, a porcentagem total dos dados no intervalo entre a média e $\pm MAD$ corresponde à 75% do total da distribuição. O MAD é muito interessante estatisticamente porque, diferentemente da média, e da variância, entre outras medidas de tendência central, é insensível a valores aberrantes ou atípicos. Por exemplo, se num conjunto de dados finito, um destes dados tende a infinito. Os valores da média e da variância, tenderão a aproximar-se de valores atípicos e não dos valores do conjunto [21].

O MAD é calculado da seguinte forma:

$$MAD = C \times med\,(\,mod\,(x_i - med(x)))\,, \tag{5.2.1}$$

sendo que mod representa o módulo, med a mediana e C uma constante que depende da distribuição. Vamos supor neste trabalho, que a fração da contagem de genes endógenos em relação a contagem dos genes alinhados,

definido como $x_i = \frac{ENDOGENOS}{ALINHADOS}$, segue uma tendência normal. Neste caso $C = 1,4826$. Portanto, a fórmula do MAD será:

$$MAD = 1,4826 \times med\left(mod\left(\frac{ENDOGENOS}{ALINHADOS} - med\left(\frac{ENDOGENOS}{ALINHADOS} \right) \right) \right).$$

Desta forma, passarão pelo segundo controle de qualidade as amostras que ficarem dentro do seguinte intervalo:

$$med\left(\frac{ENDOGENOS}{ALINHADOS} \right) - MAD \leq \frac{ENDOGENOS}{ALINHADOS} \leq med\left(\frac{ENDOGENOS}{ALINHADOS} \right) + MAD$$

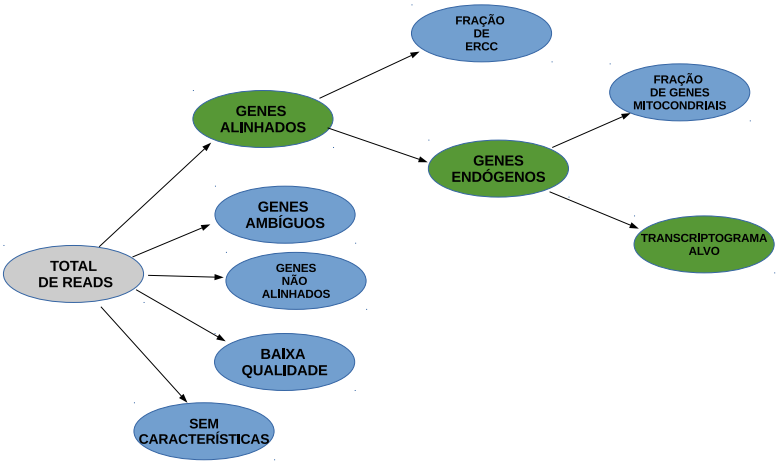

Figura 5.1: Esquema do Controle de Qualidade típicos de RNASeq

Passando por este segundo controle de qualidade, sobraram 58 células na fase G1, 59 células na fase S e 65 células na fase G2M. Para a realização da análise por transcriptograma e PCA, não será desconsideradas as amostras que não passaram pelo segundo controle de qualidade, já que, diferente do primeiro controle, estas amostras não apresentam baixa qualidade (que poderia atrapalhar na análise), mas simplesmente a confiança na classificação em alguma fase do ciclo celular é baixa. Entretanto, nas nossas análises estas amostras serão identificadas e diferenciadas das demais.

5.2.2 Normalização

O segundo passo, depois de passar as amostras por um controle de qualidade, é fazer o adequado ordenamento dos genes, aproximando genes pela função biológica que exercem. O método utilizado foi o método da função custo, desenvolvido como explicado na seção 4.1.1.

Depois de ter os dados devidamente ordenados, de acordo com as considerações explicadas anteriormente, é necessário normalizar os dados. A normalização dos dados é importante para corrigir os erros de determinadas etapas do RNASeq, como por exemplo, no momento de transformar o mRNA em cDNA ou durante os ciclos de duplicação do cDNA.

Existem muitas normalizações típicas para o tratamento dos dados obtido por RNASeq. A que iremos utilizar é a TPM(Transcript Per Million)[22]. Esta normalizaçao consiste em:

1. Primeiro dividir a contagem associada a cada gene pelo comprimento do respectivo do gene (gene length).
2. Depois somar, para cada amostra, o valor total de contagem (já divididas pelo comprimento do gene) e dividir a contagem pela soma total na amostra. Isso significa:

$$\tau_i = \frac{C_{i,a}}{l_i}\left(\sum_a \frac{C_{i,a}}{l_i}\right)^{-1} \tag{5.2.2}$$

Sendo $C_{i,a}$ o valor da contagens do gene i na amostra a, l_i o comprimento do gene i. A normalização está baseado na suposição de que genes com um comprimento maior tendem a ser mais expressos. Também está sendo normalizado pela contagem total, por que no processo de multiplicação do cDNA experimentalmente, pode ocorrer que algumas amostras seja mais multiplicadas que outras, ficando complicado a comparação entre amostras na análise.

O objetivo na realização do PCA, como foi explicado na seção 4.2 é reduzir a dimensionalidade deste conjunto de dados, em poucas dimensões que contenham a informação relevante ao sistema. Quanto mais informação estiver contido num número de reduzido de variáveis, e portanto quanto maiores as variâncias contidas nestas poucas variáveis, melhor será a representatividade do sistema. Então procurou-se aquele raio, para esta normalização, que apresentavam maiores variâncias contidas em poucas dimensões. Foi notado que escolhendo o raio em torno de 30, obtinha resultados muito bons em relação ao PCA. As figuras 5.2 e 5.3, utilizando raio 30, apresentam a variância acumulada do PCA e a variância absoluta para os dados normalizados por TPM. respectivamente.

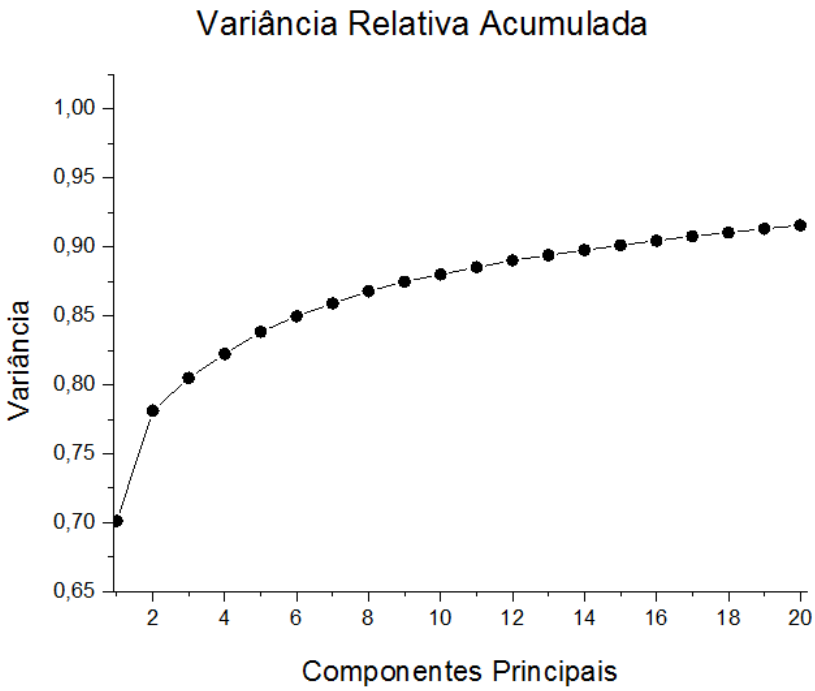

Figura 5.2: Variância Relativa Acumulada para as primeiras componentes principais

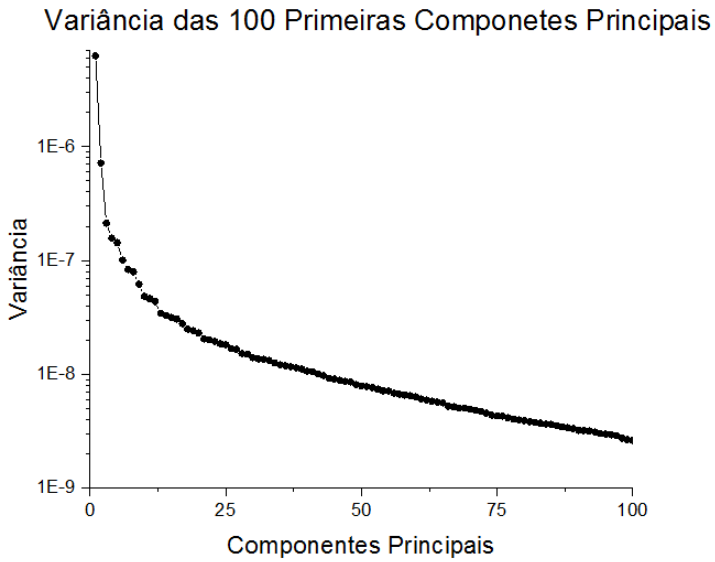

Figura 5.3: Variância absoluta das 100 primeiras componentes principais. Eixo da variância em escala logarítmica.

Analisando o gráfico 5.3 percebemos que a partir de em torno da vigésima componentes principal, a variância apresenta um comportamento típico aleatório. Isso indica que as primeiras componentes são aquelas que apresentam realmente informações relevantes para entender o sistema.

As figuras 5.4 e 5.5, demonstra o poder de reconstrução das amostras pelas componentes principais: comparamos o transcriptograma de uma amostra com o transcriptograma estimado utilizando apenas as 3 primeiras componentes principais. As figuras mostram que o transcriptograma reconstruído contorna o transcriptograma real, de modo que, as 3 principais componentes apresentam uma grande representatividade do perfil de expressão de cada amostra.

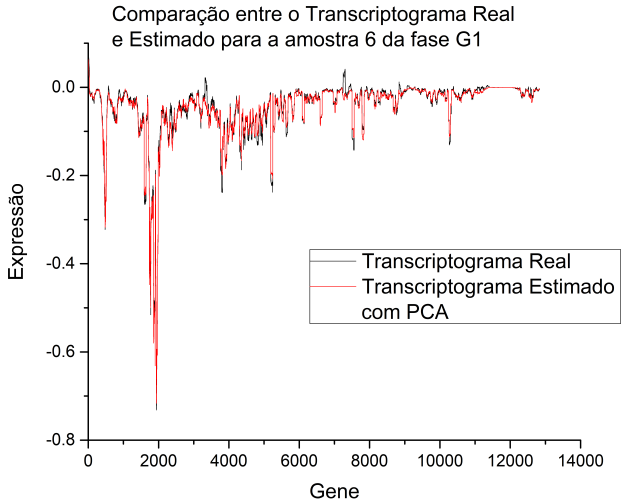

Figura 5.4: Comparação do Transcriptograma Real para a amostra 6 na fase G1 com um transcriptograma estimado utilizando as 3 primeiras componentes principais.

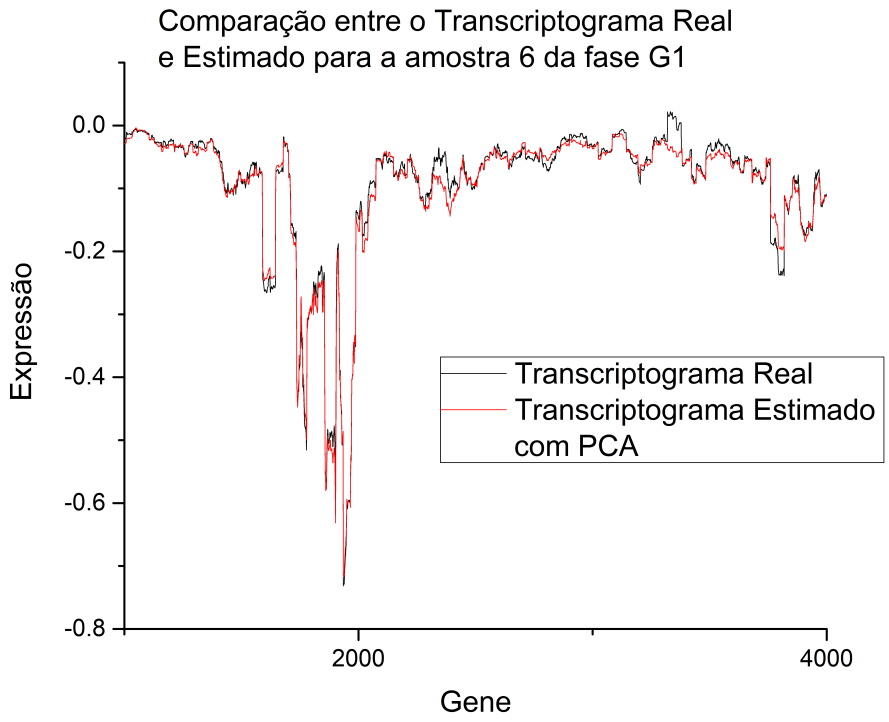

Figura 5.5: Comparação do Transcriptograma Real para a amostra 6 na fase G1 com um transcriptograma estimado utilizando as 3 primeiras componentes principais numa região dos genes 1000 a 4000.

5.2.3 Separação das Amostras e Ordenamento

Observando a figura 5.2, notamos que com a normalização TPM, para uma componente maior que 3, há uma variação em torno de 20% (mais de 80% da variação está contida nas 3 primeiras componentes), que está sendo repartida nas outras 242 componentes. Vamos supor neste trabalho, que a informação relevante para o ciclo celular, está contida nas 3 primeiras componentes. Com esta hipótese é válido normalizar os valores dos coeficientes das amostras referentes às 3 primeiras componentes. Embutido nesta hipótese, está a de que a quantidade total de *reads* não traz informação relevante, mas sim o perfil relativo de expressão de todos os genes. Isso equivale a dizer que, no espaço das principais componentes, a informação biológica está contida na direção do vetor que representa cada amostra e não no seu módulo.

O critério final, usado para validar a representação de toda a informação biológica referente ao ciclo celular nas 3 primeiras componentes, será a capacidade que estas componentes têm de discriminar as amostras na sua fase no ciclo celular (como posteriormente será mostrado que é capaz).

A normalização das componentes é feita da seguinte maneira: Seja t_i^a o valor do transcriptograma associado à i-ésima posição do ordenamento da a-ésima amostra, então:

$$t_i^a = \sum_{j=1}^{245} c_j^a p_i^j \,, \tag{5.2.3}$$

onde a soma sobre j corresponde à soma sobre as 245 componentes principais, c_j^a é o coeficiente da amostra

a na direção da j-ésima componentes principal e p_i^j representa a j-ésima componente principal, que nada mais é que um transcriptograma normalizado, de forma que:

$$\sum_{i=1}^{N} \left(p_i^j\right) = 1\,. \tag{5.2.4}$$

Utilizando apenas as 3 primeiras componentes principais, isto é, a projeção do transcriptograma de uma amostra sobre os novos eixos representados pelas componentes principais, a normalização fica:

$$\left(c_1^a\right)^2 + \left(c_2^a\right)^2 + \left(c_3^a\right)^2 = 1 \tag{5.2.5}$$

A equação 5.2.5 equivale a normalizar (em módulo 1) a projeção de cada transcriptograma sobre o espaço tridimensional gerada pelas 3 primeiras componentes principais, colocando todas as amostras sobre a esfera unitária neste espaço.

A figura 5.6, a seguir mostra a projeção das 245 amostras (incluindo as amostras que não passaram pelo segundo controle de qualidade) representadas pelas 3 primeiras componentes. Como as 2 primeiras componentes acumulam a maior parte da variação, a normalização 5.2.5 faz com que a maioria das amostras disponham sobre um círculo no plano $PC1 \times PC2$. As amostras que ficam no interior deste círculo, são aquelas que estão mais alinhadas a $PC3$.

Analisando a figura 5.6 notamos que a componente $PC1$ tem a capacidade de discriminar as amostras da fase G1 das outras. Já as amostras na fase S e G2M são separadas pela $PC3$. Com isso, obtemos regiões onde a maior parte das amostras de uma determinada fase se encontram.

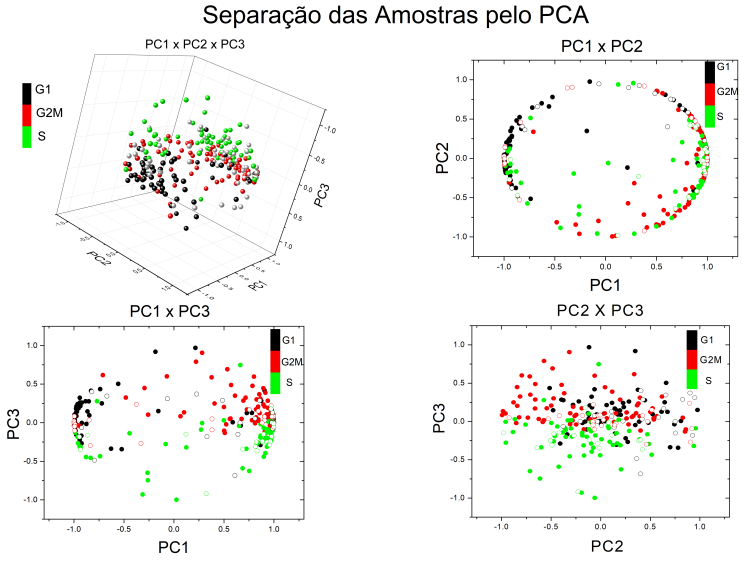

Figura 5.6: Layout da separação das amostras no ciclo celular a partir das 3 primeiras componentes principais. Os círculos vazados representam as amostras que não passaram pelo segundo controle de qualidade.

Procurando agora um ordenamento pseudo-cronológico nas amostras, seguindo a lógica biológica do ciclo celular, isto é, ao longo do ciclo, cada célula entra na fase G1, depois S e então G2M. Para tal, será feito uma primeira mudança de coordenadas, de forma a projetar este espaço tridimensional num plano bidimensional. A mudança de coordenadas foi feita da seguinte forma:

$$\theta = arctg\left(\frac{PC1}{PC2}\right) \quad ; \quad \phi = arctg\left(\frac{PC3}{\sqrt{(PC1)^2 + (PC2)^2}}\right) \tag{5.2.6}$$

Os ângulos θ e ϕ serão chamados de longitude e latitude, respectivamente, em analogia à latitude e longitude sobre o globo terrestre. O gráfico 5.7 apresenta o plano bidimensional formado por esta transformação de coordenadas. Observamos com esta mudança de coordenadas, que existe uma clara separação das amostras, existindo regiões de agrupamento das amostras nas respectivas fases do ciclo celular.

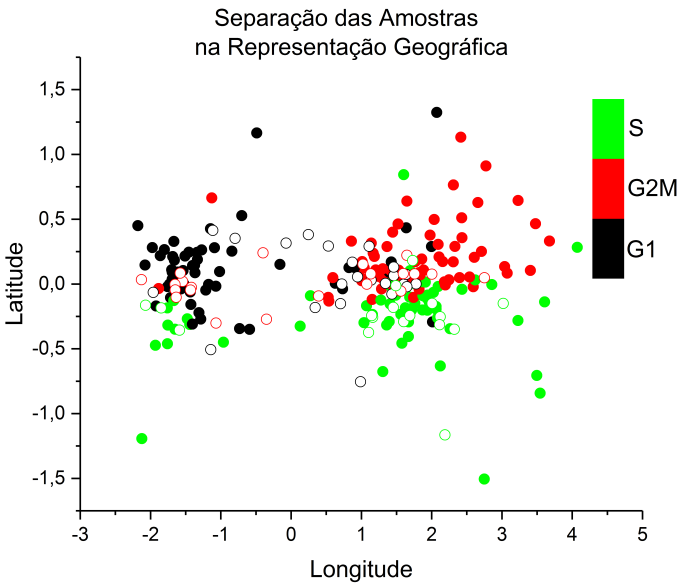

Figura 5.7: Separação das amostras no ciclo celular na nova representação geográfica das 3 primeiras componentes principais. Os círculos vazados representam as amostras que não passaram pelo segundo controle de qualidade.

Agora é necessário escolher um caminho para ordenar estas amostras, que estão claramente separadas pelas características dos seus perfis de expressão. O caminho escolhido foi de obter o ângulo, para cada amostra, entre o eixo longitude com a reta que liga a amostra ao centro do eixo de coordenadas. Isso significa, analogamente, obter o ângulo polar, em coordenadas polares. O ângulo polar ρ é calculado da seguinte forma:

$$\rho = arctg\left(\frac{LONGITUDE}{LATITUDE}\right) \tag{5.2.7}$$

O ordenamento das amostras foi realizado pela ordem crescentes dos valores do ângulo ρ. A figura 5.8 apresenta este ordenamento, em função dos valores de PC1, PC2 e PC3. Analisando as figuras 5.8.a e 5.8.c representando este ordenamento em relação a PC1 e PC3 existe um caminho bem definido que as amostras seguem no transcorrer do ciclo celular, em especial representando em função da PC3. Já o ordenamento em função do valor da PC2 não apresenta nenhum padrão de caminho aparente. Isso pode ser explicado pelas

características de cada componente principal. Diferentemente da PC2, as componentes PC1 e PC3, pelo gráfico 5.6 têm o poder de discriminar características das amostras em cada fase do ciclo celular, separando-as.

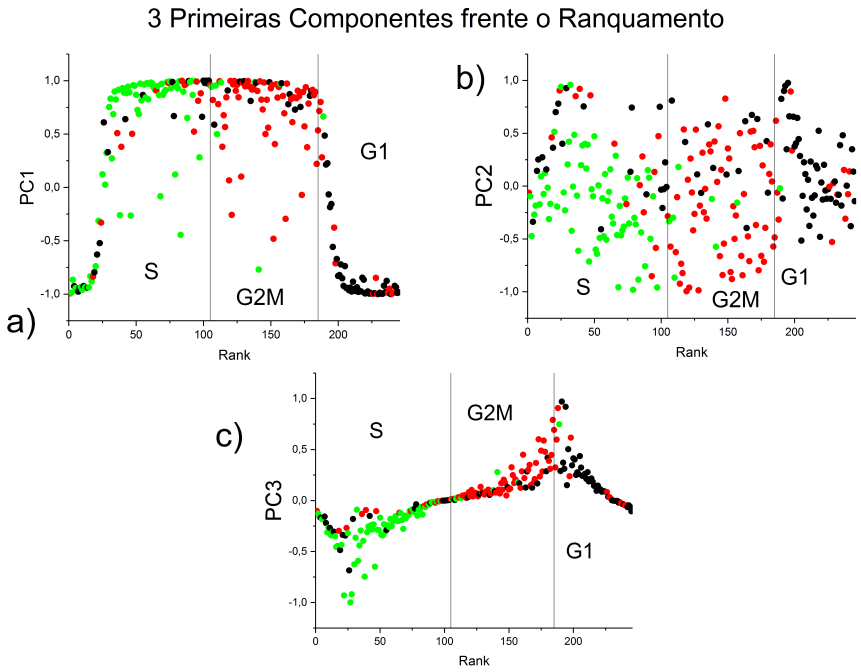

Figura 5.8: Ordenamento das amostras passando pelo ciclo celular, referentes aos valores das 3 primeiras componentes principais. (a) Ranqueamento referente a PC1. (b) Ranqueamento a PC2 e (c) ranqueamento referente a PC3.

Agora vamos graficar intensidades coeficientes das componentes principais, bem como a intensidade de expressão de genes ou conjuntos de genes em função da ordem proposta pelo ordenamento acima explicado. Para tanto, quando necessário, iremos suavizar as curvas da figura 5.8. Para isso, será utilizado o filtro de Savitzky–Golay[23], que é um método baseado no cálculo de regressão polinomial local. Para fazer o método é utilizado $K+1$ pontos igualmente espaçados em uma curva, aproximando estes pontos por um polinômio (neste trabalho será utilizado um polinômio de grau 2), resultando em uma curva parecida com a original, entretanto suavizada. Utilizaremos 20 pontos para fazer o polinômio. A curva aproximada preserva as característica da original, como por exemplo os máximos e mínimos relativos. Para realizar computacionalmente este filtro, utilizaremos a ferramenta OriginLab, na aba de técnicas de processamento de sinal.

Comparando as figuras 5.8 e 5.9, notamos que, referente ao ordenamento e os valores de PC1 e PC3, com a suavização, observamos um claro caminho (5.9.a e 5.9.c) , já observado com a curva original (5.8.a e 5.8.c). Entretanto, suavizando a distribuição do ordenamento das amostras com os valores de PC2 (5.8.b), obtemos um caminho claro, de quais são os valores das amostras, referentes ao PC2, passando ao longo do ciclo celular (5.9.c).

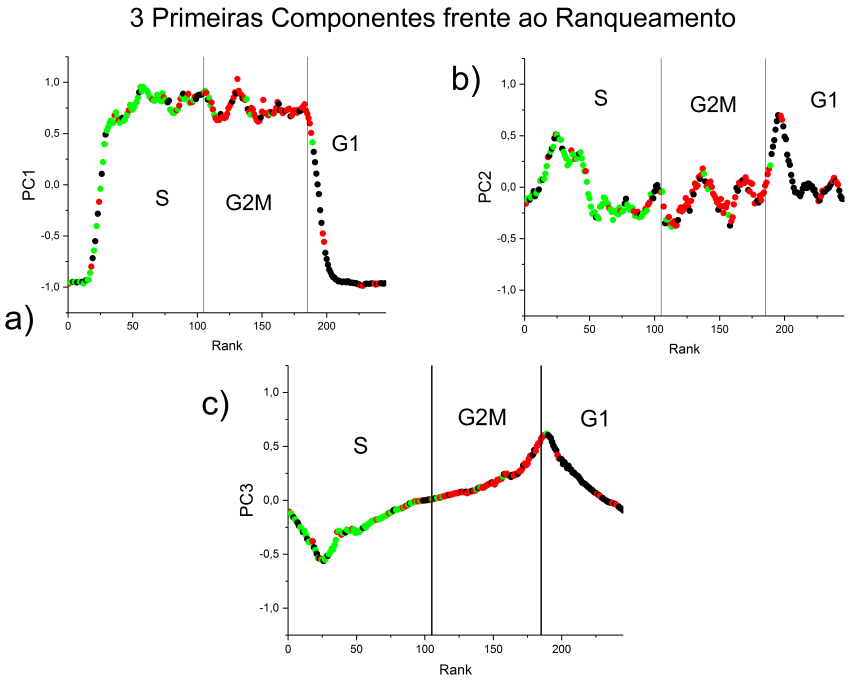

Figura 5.9: Ordenamento das amostras passando pelo ciclo celular, referentes aos valores das 3 primeiras componentes principais passando pelo filtro de Savitzky–Golay. (a) É referente a PC1. (b) Referente a PC2 e (c) é referente a PC3.

5.2.4 Validação Biológica

O ordenamento das amostras, que sugere a evolução da célula ao longo do ciclo celular, precisa ser validado biologicamente. Para isso, usaremos as informações sobre o complexo ciclina-CDK, como visto na seção 2.2. Para avaliar a expressão do complexo ciclina-CDK, foi multiplicado o valor da expressão da ciclina com valor da expressão da CDK, na respectiva amostra. Isso é feito porque o funcionamento do complexo depende dos dois agentes e portanto, amostras que tiverem um valor alto de expressão de apenas um deles, não representa necessariamente que o complexo está em funcionamento. A figura 5.10 apresenta a expressão de alguns complexos ciclina-CDK.

Analisando a figura 5.10, observamos que: A ciclina D (5.10.d), que começa a sua expressão em G1 e termina na mitose, sendo expressa ao longo de todo o ciclo, faz sentido com o valor da expressão do nosso ordenamento, sendo expressa ao longo de todo o ciclo, mas com uma queda no final de G2M, representando a mitose. O complexo Ciclina A e CDK1 (5.10.a) que está associado à passagem da célula para a mitose, está bem representada no nosso ordenamento, sendo mais expressa na G2M, assim como o complexo Ciclina B e CDK1 (5.10.c). A ciclina E associada a CDK2 (5.10.b) tem sua maior expressão na passagem da célula de G1 para S, característica encontrada no ordenamento.

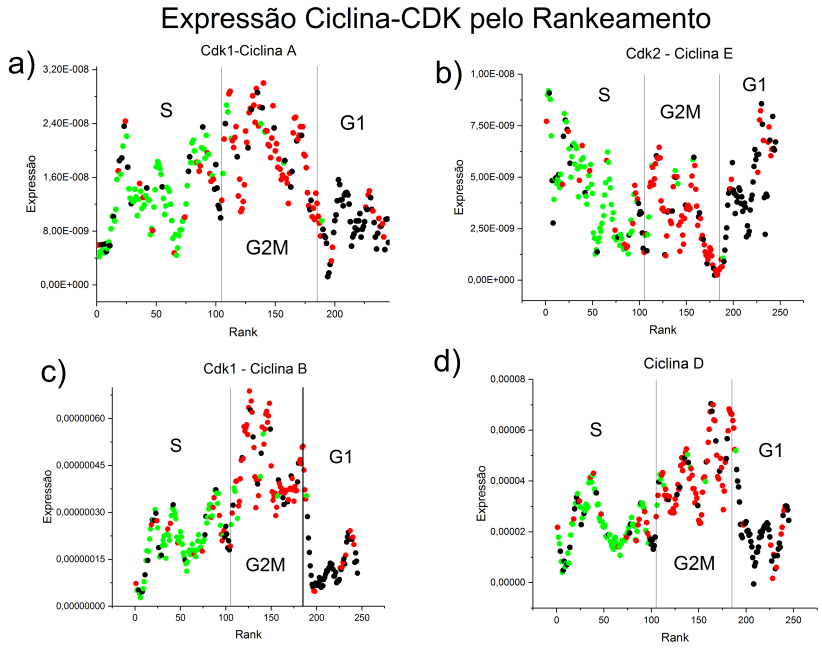

Figura 5.10: Valor de expressão de alguns complexos Ciclina-CDK ao longo do nosso ordenamento, passando pelo ciclo celular. (a) Refere-se ao complexo Ciclina A e CDK 1. (b) Ao complexo Ciclina E e CDK 2. (c) Complexo Ciclina B e CDK 1. (d) A Ciclina D.

Capítulo 6

Conclusão

Nestes trabalho, utilizando dados de expressão gênica por RNASeq das células únicas de *Mus Musculus*, foi possível, pelo método do PCA, separar as amostras de acordo com a fase no ciclo celular, utilizando as 3 primeiras componentes principais. Com a separação obtida, também foi possível criar um ordenamento destas amostras, representando a passagem da célula pelo ciclo celular. Portanto, o PCA é método interessante para se obter informações deste tipo de sistemas com muitas variáveis e funções internas complexas.

O transcriptograma têm um papel muito importante na análise dos dados. Como já foi dito anteriormente, determinadas funções biológicas são exercidas por muitos genes diferentes, então fazer médias sobre os vizinhos ordenados, ou seja, fazer uma média da expressão de genes que colaboram nessas funções é uma forma válida de não perder as características principais do sistema, ao mesmo tempo que melhora a razão do sinal-ruído.

Encontramos uma ordem pseudo cronológica dos genes que foi validada utilizando-se genes marcadores de determinadas etapas do ciclo celular. Isso mostra uma coerência entre a ordem criada com as amostras (que são "fotografias"das células em determinado momento do ciclo celular), com a realidade metabólica do ciclo celular.

A continuação deste trabalho será aperfeiçoar o ordenamento das amostras, encontrando ao longo do ciclo, rotas que melhor representem os padrões cíclicos bem fundamentados de determinados grupos de genes. Depois de encontrado o melhor ordenamento possível das amostras, será possível, por exemplo, gerar o perfil de expressão esperado deste ordenamento, verificando as variações das funções biológicas, além de analisar quais genes aumentam ou diminuem na medida que a célula passa pelo ciclo celular.

Referências Bibliográficas

[1] Luis Carlos Junqueira and José Carneiro. *Biologia Celular e Molecular*. Guanabara Koogan, Rio de Janeiro, 2012.

[2] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

[3] Imagem modificada de 'Control del Ciclo Celular' figura 2 de openstax college. biologia ((cy by 3.0)). modificação do trabalho originial de wikimama. https://cnx.org/contents/GFy_h8cu@9.87:abji7vNQ@6/Control-of-the-Cell-Cycle.

[4] Frederick Sanger, S Nicklen, and AR Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1977.

[5] AM Maxam and M Gilbert. A new method for sequencing. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977.

[6] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome.. *Nature*, 431(7011):931–945, 2004.

[7] Z Wang, M Gerstein, and M Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[8] Sandrine Imbeaud, Esther Graudens, Virginie Boulanger, Xavier Barlet, Patrick Zaborski, Eric Eveno, Odilo Mueller, Andreas Schroeder, and Charles Auffray. Towards standardization of rna quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Research*, 33(6):e56, 2005.

[9] Eric Kawashima, Laurent Farinelli, and Pascal Mayer. Patent: Method of nucleic acid amplification, 2005.

[10] A Valouev, J Ichikawa, T Tonthat, J Stuart, S Ranade, H Peckham, K Zeng, JA Malek, G Costa, K McKernan, A Sidow, A Fire, and SP Johnson . A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7):1051–1063, 2008.

[11] A Elizabeth and Charlie L Holcomb . Next-generation hla sequencing using the 454 gs flx system. *Methods in Molecular Biology*, 1034:197–219, 2013.

[12] SRM da Silva, GC Perrone, and RMC Almeida. Transcriptograms: Reproducibility enhancement and differential expression of non predefined functional gene sets in human genome. *BMC Genomics*, 15:1181, 2014.

[13] SRM da Silva. A eficiência do transcriptograma. *Master thesis - Instituto de Física, Universidade Federal do Rio Grande do Sul*, 2013.

[14] A Franceschini, D Szklarczyk, S Frankild, M Simonovic M Kuhn, A Roth, P Minguez J Lin, P Bork, C von Mering, and LJ Jensen . String v9.1: redes de interação proteína-proteína, com maior cobertura e integração. *Nucleic Acids Research*, 2013.

[15] J L Rybarczyk-Filho, M A A Castro, J C F Moreira Dalmolin RJ, L G Brunnet, and RMC de Almeida . Rumo a um transcriptograma genômico: o caso saccharomyces cerevisiae. *Nucleic Acids Research*, 39:3005–3016.

[16] Nicholas Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949. PMID: 18139350.

[17] M Ashburner and et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[18] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[19] TW Anderosn. *Introduction to multivariate statistical analysis*. Wiley-Interscience, New York, 1958.

[20] Jonathan Shlens. A tutorial on principal component analysis. https://arxiv.org/pdf/1404.1100.pdf, 2014.

[21] C Leys, C Ley, O Klein, P Bernard, and L Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[22] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):438–500, 2010.

[23] Abraham Savitzky Abraham and Marcel J E Golay. Smoothing and differentiation of data by simplified least squares procedures. . *Analytical Chemistry*, 36:1627–1639, 1964.