Daniel Matos Montezano

CONTRIBUIÇÕES PARA MODELAGEM E PREDIÇÃO DO METABOLISMO DE BACTÉRIAS EXPOSTAS À AÇÃO DE ANTIBIÓTICOS

Tese submetida ao Programa de Pós-Graduação
em Engenharia Elétrica para a obtenção
do Grau de Doutor em Engenharia Elétrica.
Orientador
Universidade Federal de Santa Catarina:
Prof. José Carlos Moreira Bermudez, Ph.D.

Florianópolis

2016

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Montezano, Daniel Matos
Contribuições para modelagem e predição do metabolismo de bactérias expostas à ação de antibióticos / Daniel Matos Montezano ; orientador, José Carlos Moreira Bermudez - Florianópolis, SC, 2016.
236 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia Elétrica.

Inclui referências

1. Engenharia Elétrica. 2. Biologia computacional. 3. Modelagem de metabolismo. 4. Pesquisa de antibióticos. 5. Tuberculose. I. Bermudez, José Carlos Moreira. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Engenharia Elétrica. III. Título.

Daniel Matos Montezano

CONTRIBUIÇÕES PARA MODELAGEM E PREDIÇÃO DO METABOLISMO DE BACTÉRIAS EXPOSTAS À AÇÃO DE ANTIBIÓTICOS

Esta Tese foi julgada adequada para a obtenção do Título de "Doutor em Engenharia Elétrica", Área de concentração Comunicações e Processamento de Sinais, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica.

Florianópolis, 23 de agosto 2016.

Prof. Marcelo Lobo Heldwein, Dr. sc ETH
Coordenador
Universidade Federal de Santa Catarina

Banca Examinadora:

Prof. José Carlos Moreira Bermudez, Ph.D.
Orientador
Universidade Federal de Santa Catarina

Prof. Vítor Heloiz Nascimento, Ph.D.
Universidade de São Paulo

Prof. André Fujita, Dr.
Universidade de São Paulo

Prof[a]. Daniela Ota Hisayasu Suzuki, Dr[a].
Universidade Federal de Santa Catarina

Prof. Carlos Aurélio Faria da Rocha, Dr.
Universidade Federal de Santa Catarina

Prof. Márcio Holsbach Costa, Dr.
Universidade Federal de Santa Catarina

Prof. Jefferson Luiz Brum Marques, Ph.D.
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais.

AGRADECIMENTOS

Agradeço especialmente aos meus pais pela dádiva da vida.

A todos os familiares e amigos que acompanharam de perto ou de longe o meu trabalho durante os anos de pesquisa como doutorando.

Aos colegas e professores do LPDS (*Laboratório de Pesquisas em Processamento Digital de Sinais*) por proporcionarem um ambiente agradável e propício ao meu desenvolvimento como pesquisador.

Ao Prof. Luiz Eduardo Bermudez da Oregon State University pela oportunidade de trabalhar em um tema ao mesmo tempo cativante e desafiador.

Aos funcionários da UFSC, Wilson Silva Costa e Marcelo Manoel Siqueira.

Ao CNPq pela oportunidade, incentivo e suporte financeiro (Processo 141.869/2011-9 da chamada GM/GD - Cotas do Programa de Pós-graduação), sem o qual o trabalho não teria sido possível.

Finalmente, agradeço de modo especial ao meu orientador Prof. José Carlos Moreira Bermudez por sua inestimável ajuda, paciência e compreensão, tanto com as minhas idiossincrasias quanto com minhas dificuldades.

Whatever the cause may be of each slight difference in the offspring from their parents – and a cause for each must exist – it is the steady accumulation, through natural selection, of such differences, when beneficial to the individual, that gives rise to all the more important modifications of structure, by which the innumerable beings on the face of this earth are enabled to struggle with each other, and the best adapted to survive.
Charles Darwin, 1859

According to Darwin's Origin of Species, "it is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself."
Leon C. Megginson, 1963

RESUMO

Neste trabalho estudou-se o metabolismo do *Mycobacterium tuberculosis* com ferramentas computacionais. Compreender a fisiologia e o metabolismo dos organismos patogênicos é de primordial importância na pesquisa e desenvolvimento de antibióticos, pois a compreensão do mecanismo de ação de uma nova droga passa pela compreensão de como a exposição do organismo ao antibiótico afeta e altera o seu metabolismo. Esse conhecimento se torna ainda mais importante em situações em que o organismo desenvolve resistência aos compostos. É sabido que o *Mycobacterium tuberculosis* possui um grande poder de adaptação e capacidade de desenvolver resistência a muitos dos compostos atualmente utilizados para o tratamento da tuberculose. As linhagens resistentes adaptam-se e desenvolvem a capacidade de sobreviver mesmo em presença de um composto a que antes eram susceptíveis. A sobrevivência do organismo está associada diretamente com a capacidade de manter o seu metabolismo em um estado funcional após exposto ao composto bactericida. A sobrevivência das cepas resistentes, portanto, depende de um uso diferenciado dos caminhos metabólicos. Estudar o metabolismo permite compreender a gama de variações metabólicas executadas pelo organismo para manter a sua sobrevivência. O estudo computacional do metabolismo em escala genômica apresenta entretanto diversos desafios. Três principais problemas são (1) o reduzido número de amostras em grande parte dos experimentos de bancada, devido ao alto custo da implementação e complexidade dos métodos de obtenção de dados (2) a inerente alta dimensionalidade dos dados e (3) como incorporar conhecimento biológico existente em modelos computacionais. Este trabalho busca soluções para contornar esses três problemas, com um estudo sobre geração de dados sintéticos, técnicas de estimação, e propostas de modelos com dimensionalidade reduzida e uso de informação biológica *a priori* em modelos computacionais. Os resultados mostram que a aplicação de técnicas computacionais para estudo de metabolismo de organismos expostos à ação de antibióticos é de fundamental importância na identificação de caminhos metabólicos de sobrevivência, além de produzir uma maior compreensão do mecanismo de ação de compostos antibióticos sobre um organismo patogênico.

Palavras-chave: Metabolismo. Biologia Computacional. Antibiótico.

ABSTRACT

In this work the metabolism of the *Mycobacterium tuberculosis* (Mtb) was studied with computational biology tools. To understand the physiology and metabolism of pathogenic organisms is of paramount importance in antibiotics R&D, because in order to understand the mechanism of action of a new drug it is necessary to understand how exposing an organism to the antibiotics affects and alters its metabolism. Even more important is this knowledge in situations when the organism may develop resistance to the compound. It is known that the *Mycobacterium tuberculosis* is highly adaptable and is able to develop resistance to a great number of the anti-tubercular compounds used for tuberculosis (TB) treatment today. The resistant strains adapt and develop the ability to survive even in the presence of a bactericidal compound for which they were once susceptible. The survival of the organism is directly linked to its ability to keep metabolism in a functional state after drug exposure. Therefore, the survival of Mtb resistant strains hinges on being able to use its metabolic mechanisms in a different way. By studying metabolism one may be able to understand the large spectrum of metabolic variations performed by the organism in order to remain alive. A computational study of metabolism in a genomic-scale however presents a number of challenges. Three main difficulties that arise are: (1) the small number of samples available in the majority of wetlab experiments, due to high cost and complexity of methods used for experimental data gathering, (2) the high dimensionality of the data and (3) how to incorporate available biological knowledge in computational models. This work has searched for solutions to overcome these three problems with a study on synthetic data generation, estimation tools, models with reduced dimensionality, and with inclusion of *a priori* biological knowledge. The results show that the use of computational techniques to study the metabolism of organisms exposed to antibiotics is of fundamental importance for identification of metabolic pathways used for survival and also for producing a better understanding of the mechanism of action of antibiotic compounds on a pathogenic organism.

Keywords: Metabolism. Computational Biology. Antibiotics.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Neste trabalho são estudadas algumas técnicas computacionais para modelagem do metabolismo do *Mycobacterium tuberculosis* e de atividade de reações e caminhos metabólicos.

Iniciaremos a apresentação com uma breve discussão sobre modelos computacionais na área da biologia e as dificuldades associadas, destacando os diversos tipos de dados experimentais que podem ser utilizados na estimação e aprimoramento de tais modelos.

Apresentamos também alguns dos problemas associados às técnicas computacionais na área de biologia, e a necessidade de aprimoramento dessas técnicas e dos modelos que necessitam ser desenvolvidos.

O único organismo tratado neste trabalho é o *Mycobacterium tuberculosis*, e o seu metabolismo é estudado a partir de uma visão computacional. Por esse motivo, no final deste capítulo introdutório discutiremos o aspecto do metabolismo desse organismo que motivou as técnicas desenvolvidas no restante do trabalho, o metabolismo de sobrevivência. O metabolismo de sobrevivência do *Mycobacterium tuberculosis* é resultante da sua alta capacidade de adaptação, que permite modificar a sua configuração metabólica em resposta a variados estresses fisiológicos como depleção de nutrientes e exposição a antibióticos. Essa alta capacidade de adaptação garante a sua sobrevivência e reprodução em condições ambientais mais hostis, bem com o desenvolvimento de linhagens mais resistentes.

1.1 MODELOS BIOLÓGICOS COMPUTACIONAIS

Neste trabalho apresentamos diversas técnicas computacionais de predição para atividade de caminhos metabólicos. O objetivo ao apresentar essas técnicas é fornecer diretrizes e auxiliar na busca de alvos moleculares para novos compostos antibióticos no combate à tuberculose. A pesquisa por novos e melhores modelos preditivos para o metabolismo do *Mycobacterium tuberculosis* (e de outros organismos patogênicos) justifica-se:

- no caso do desenvolvimento de novos compostos antibióticos e identificação de novos alvos genéticos/proteicos/metabólicos;

- para interpretação de dados experimentais que levem à elucidação e criação de hipóteses sobre o funcionamento do sistema me-

tabólico;

- para aprimorar a predição de características fenotípicas do organismo, condição essencial não apenas no processo de desenvolvimento de compostos antibióticos, mas na própria necessidade de compreensão da fisiologia dos organismos procariotas em geral e suas interações com organismos hospedeiros.

A biologia, tradicionalmente de caráter reducionista, hoje necessita ser estudada e compreendida de forma global. É necessário estudar as células de forma integrada, como sistemas complexos que são, ao invés de apenas analisar suas partes separadamente, na esperança de que as conclusões individuais necessariamente levem a conclusões sobre o sistema global.

É consenso entre pesquisadores da área que métodos computacionais são necessários para estudar os sistemas biológicos de forma integrada (PALSSON, 2000). Não é suficiente a análise da função de apenas um gene ou de uma única proteína em um número reduzido de condições experimentais. É essencial que a função de diversos genes e proteínas possa ser analisada de forma conjunta, e no mais variado conjunto de condições possível. O primeiro passo nessa direção já foi dado com o surgimento de técnicas experimentais que produzem uma visão em escala genômica ou quase genômica da expressão genética e proteica.

Tradicionalmente, técnicas de perturbação molecular, imageamento e homologia permitiram identificar e compreender de forma individual a função de uma grande quantidade de genes e proteínas, antes da existência de técnicas experimentais de alta densidade (*high throughput*). Mesmo uma grande parte das redes genéticas de regulação e de reações metabólicas conhecidas hoje pôde ser elucidada com experimentos monogene ou monoproteína em escala reduzida. Em geral estes experimentos foram realizados com organismos unicelulares como bactérias e fungos e seus resultados transpostos com sucesso para os organismos multicelulares, e mesmo para as células do corpo humano.

Todavia, esse foco de pesquisa sobre a identidade de genes e proteínas individuais foi transladado para uma visão genômica quando experimentos que permitem medir a atividade de milhares de genes de um organismo ao mesmo tempo se tornaram disponíveis (TÖZEREN; BYERS, 2004). Como exemplo, e especificamente para o organismo *Mycobacterium tuberculosis*, técnicas experimentais de medição de expressão genética atuais podem medir para uma mesma condição experimental a atividade dos aproximadamente 4000 genes identificados

no genoma dessa bactéria, bem como quantificar uma enorme quantidade de RNAs não-funcionais (i.e. que não codificam proteínas) e que são essenciais nos processos de comunicação intracelular e de regulação (MATTICK, 2003). Outras técnicas de alta densidade como proteômica e metabolômica completam o conjunto destas ferramentas biológicas de larga-escala (*large-scale biology*) para produção de dados experimentais.

Entretanto, é importante notar que, em geral, não apenas mais experimentos precisam ser realizados, mas urge melhorar as técnicas computacionais que permitam utilizar os dados já existentes. É importante combinar esses dados em modelos mais compreensivos com o intuito de extrair informações e conhecimento que apenas surgem quando dados de diferentes fontes são utilizados. Um modelo biológico mais compreensivo pode predizer melhor os estados fenotípicos utilizando a informação combinada que existe em conjuntos de dados correlacionados (YIZHAK et al., 2010). Porém, *como* incorporar e combinar essas informações nos modelos é o que ainda representa um dos grandes desafios da biologia computacional.

Como observou-se acima, não é mais possível analisar essas grandes quantidades de dados sem técnicas computacionais de alto desempenho. A análise dos dados biológicos de alta densidade atuais necessariamente passa por procedimentos matemáticos e computacionais que envolvem técnicas de processamento digital de sinais, de reconhecimento de padrões, de aprendizado de máquinas (*machine learning*), entre outros, e portanto definem um quadro no qual a multidisciplinaridade é um fator chave. A interação entre biólogos, geneticistas, engenheiros, estatísticos e cientistas da computação é um elemento essencial para o sucesso de projetos de pesquisa atuais na área da biologia.

O desenvolvimento de um modelo inicia com o planejamento de experimentos e a obtenção dos dados experimentais. Com o significativo avanço e aprimoramento dos últimos anos nas técnicas experimentais, não apenas um maior volume de dados pôde ser obtido, mas também uma maior variedade no tipo de informação (e.g. dados de proteômica, micromatrizes de DNA, metabolômica, construção de redes biológicas, identificação da localização e da função dos genes e suas interações a nível molecular, ação de enzimas catalisadoras, variações em fluxos metabólicos, etc.). A partir desses novos conjuntos de dados, os organismos, tanto procarióticos quanto eucarióticos, podem agora ser observados em sua extrema complexidade, de forma mais completa e sob diferentes pontos de vista, estruturais e funcionais.

Na Fig. 1, mostramos de forma simplificada o fluxo da informação no desenvolvimento de um modelo biológico computacional. Apresen-

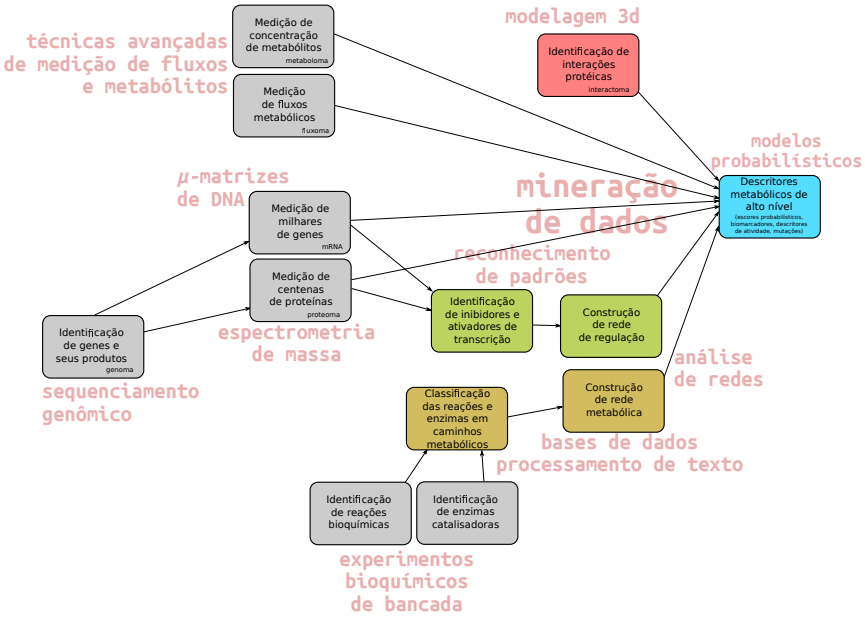

Figura 1 – Diagrama simplificado do fluxo de informação no desenvolvimento de um modelo biológico computacional. *Fonte: Elaborada pelo autor.*

tamos nessa figura diversas técnicas de obtenção de dados biológicos, os respectivos conjuntos de dados obtidos e como tais dados são transformados em informação e conhecimento acerca dos organismos. Algumas técnicas experimentais, embora apresentando contínuo aprimoramento ao longo da última década, já estão plenamente estabelecidas. As técnicas de sequenciamento genômico e de medição de expressão genética, bem como a identificação de conteúdo proteico com espectrometria de massa (produzindo um quadro geral do conjunto de proteínas existente em um determinado instante temporal e condição experimental dentro da célula) são úteis no que permitem observar de forma bastante clara as respostas da célula a estímulos externos, a ativação e desativação de diferentes programas celulares após alterações ambientais, com subsequente renovação (degradação e biosíntese) no proteoma e, finalmente, alterações no metabolismo realizadas pelo complexo celular. Outras tecnologias, embora igualmente importantes, ainda sofrem com a necessidade de melhorias tecnológicas, como a medição do metaboloma (ROESSNER; BOWNE, 2009) e do fluxoma (WITTMAN, 2007).

Todas as técnicas experimentais indicadas acima e mostradas na

figura são voltadas para medições completas da célula, isto é, medidas em escala genômica. Por exemplo, um experimento de micromatriz de DNA busca, de maneira geral, medir os níveis de expressão genética de *todos* os mRNAs expressos pela célula, i.e. o seu transcriptoma como um todo. No caso da medição do proteoma, embora com alto grau de precisão, as técnicas são limitadas à identificação de proteínas que já foram catalogadas, pois existe necessidade de comparar os resultados dos fragmentos com possíveis proteínas-alvo já conhecidas. Experimentos de metabolômica também têm o objetivo de medir a concentração de *todos* os metabólitos presentes na célula, assim como o fluxoma pretende fornecer medidas de fluxo metabólico para *todas* as reações bioquímicas que compõem os caminhos metabólicos do organismo, embora limitações nas técnicas ainda impossibilitem estes objetivos.

É importante notar que uma grande parte do conhecimento que se tem hoje da fisiologia dos organismos antecede a existência das técnicas em escala genômica. Essas técnicas, obtidas a partir de experimentos de bancada em menor escala, ainda são necessárias, por permitirem isolar partes do sistema celular e observar detalhes que podem permanecer mascarados sob a grande quantidade de dados gerada nos experimentos de alta densidade modernos. É certo que tais experimentos de caráter reducionista não permitem compreender o sistema celular como um todo, porém são necessários na compreensão dos detalhes. Uma analogia com um circuito eletrônico pode ser feita, em que a compreensão de um sub-circuito transistorizado que é parte de um circuito maior não permite identificar a função do circuito maior do qual ele faz parte.

Assim que, a importância de tais experimentos não pode ser subestimada. Deve-se a esses experimentos em menor escala a origem das atuais bases de dados biológicos. Essas bases de dados compreendem uma enorme gama de informações, como o funcionamento metabólico dos organismos, ou dados detalhados sobre as complexas redes de interações regulatórias e sinalização celular.

No caso de bases de dados com informações metabólicas, é possível obter conhecimento detalhado sobre reações metabólicas, sua estequiometria e enzimas associadas, os diversos papéis dessas enzimas no ambiente intracelular e sua interação com os diversos metabólitos processados dentro de cada caminho metabólico (conjunto de reações metabólicas atuando de forma coordenada, em sequência ou paralelamente, em direção a um objetivo metabólico específico). Já no caso das redes de sinalização e regulação, os dados armazenados nestes repositórios compreendem os complexos sistemas de controle celular como,

por exemplo, o que atua na organização do processo de divisão celular (reprodução mitótica para organismos procariontes). Com as informações disponíveis nessas bases de dados tem-se acesso aos diversos mecanismos de segurança e controle que ocorrem no interior da célula e garantem robustez e baixíssima taxa de erro em um processo tão essencial e delicado. É possível identificar quais proteínas ou RNA de sinalização são responsáveis pela ativação e desativação da síntese de proteínas em cada momento celular, e mesmo conhecer o caminho de processamento da informação que permite iniciar processos irreversíveis como a apoptose ou morte celular programada (PCD - *programmed cell death*). Todo esse conhecimento foi inicialmente derivado de experimentos de bancada que focavam em genes, proteínas e caminhos celulares individuais. A existência de dados de alta densidade agora permite colocar sob um mesmo foco e de forma integrada essas informações originalmente obtidas de forma separada.

Nas medidas em escala genômica é inevitável que haja redundância. Os organismos vivos são redundantes por natureza, pois é a redundância interna do sistema e a consequente robustez que permite muitas vezes a sobrevivência. Sendo assim, ao trabalharmos com medidas em escala genômica, são necessários algoritmos que possam selecionar os descritores de maior relevância, diminuir a dimensionalidade do problema, e permitir um processamento de dados mais ágil. Um número bastante significativo de técnicas de processamento de dados, estimação e classificação, (PCA, clusterização, redes neurais, etc.) e as suas inúmeras variações e aprimoramentos (IRIGOIEN; VIVES; ARENAS, 2011; THEODORIDIS; KOUTROUMBAS, 2008) está disponível ao investigador que necessita extrair informação de dados biológicos. Neste trabalho, analisamos o uso de métodos que mapeiam dados biológicos para diferentes representações e que permitem a redução da sua dimensionalidade (Capítulo 3 e 4) mantendo ainda uma quantidade de informação suficiente para procedimentos subsequentes de regressão e classificação.

É provável que modelos biológicos se beneficiem com o uso de técnicas bayesianas, as quais permitem incorporar no modelo o conhecimento *a priori* disponível, conforme indicado acima. Por exemplo, correlação entre a atividade de vários genes e proteínas disponíveis em redes de regulações genéticas, caracterização de ruído biológico em experimentos de micromatrizes de DNA, utilização da estrutura de caminhos metabólicos, interações proteicas (interactoma) (KUNDROTAS; VAKSER, 2010), entre outros, todos podem ser modelados e incorporados na forma de distribuições *a priori*. Este conhecimento do com-

portamento celular *in vitro* e *in vivo*, obtido com os experimentos de bancada individuais já discutidos e aprimorados a cada dia com novas inferências e validações feitas a partir de dados de alta densidade, estão armazenados em grandes bases de dados, em sua maioria disponíveis publicamente, tópico que discutimos brevemente a seguir.

1.2 BASES DE DADOS BIOLÓGICOS

Não é possível falar de modelagem de sistemas biológicos e técnicas de biologia computacional sem mencionar o papel essencial das bases de dados, as quais foram mencionadas brevemente acima. Diferentemente de outras áreas em que dados experimentais são obtidos para estimar modelos, na área biológica uma grande quantidade de informação *a priori* já está disponível, embora muitos modelos continuem a ser essencialmente *data-driven*. A quantidade e a qualidade das bases de dados biológicos aumentaram significativamente na última década devido aos investimentos feitos nessa área. Tais investimentos foram essenciais após o vertiginoso aumento na quantidade de dados, a ubiqüidade da informação online e o aprimoramento das técnicas computacionais de processamento de dados. Devido ao escopo deste trabalho, não é possível nomear aqui todos os tipos de bases de dados, porém as bases de dados primárias essencialmente organizam e disponibilizam dados biológicos como (LESK, 2007):

- sequências genômicas, incluindo genomas completos;
- estruturas proteicas primárias (seqüências de amino-ácidos);
- estruturas secundárias, terciárias e quaternárias de proteínas (imagens cristalográficas e modelos tridimensionais computacionais);
- funções de proteínas (e.g. enzimas e fatores de transcrição);
- estruturas moleculares de compostos (e.g. metabólitos);
- padrões de expressão genéticos, proteicos, metabólicos;
- redes de caminhos metabólicos, de interação, além de controle e regulação.

Um esforço mundial muito grande existe na tentativa de integrar cada vez mais as diferentes bases de dados, para eliminar redundâncias e erros. Algumas das bases de dados mais importantes que podemos citar

são a GEO (*Gene Expression Omnibus* - NCBI) (EDGAR; DOMRACHEV; LASH, 2002), com experimentos de micromatrizes de DNA, a Uniprot (*UniProt Consortium*) (APWEILER et al., 2004), com informações detalhadas sobre o proteoma de uma multitude de organismos, KEGG Atlas (Kanehisa Labs, 2015), com uma gama muito grande de diferentes tipos de dados integrados em redes correlacionando estrutura e função, além de diversas outras bases especializadas e dedicadas a organismos específicos. No caso do *Mycobacterium tuberculosis*, algumas das bases mais importantes, que reúnem não apenas dados sobre o organismo, mas também sobre a doença, são o TBDB (TB Database, 2015b) (*Tuberculosis Database*) e a Tuberculist (Swiss Institute of Bioinformatics and École Polytechnique Fédérale de Lausanne, 2013) (*Mycobacterium tuberculosis database*).

Todavia, o esforço de organização da informação biológica em bases de dados não seria útil sem a disponibilidade de ferramentas que permitam realizar a mineração dos dados (*data mining*) e a busca das informações desejadas (*information retrieval*). Atualmente, todos os tipos de dados biológicos caem sob a rótulo de dados grandes (*big data*), e técnicas computacionais de processamento, inferência e visualização cada vez melhores e mais rápidas são necessárias para lidar com tais conjuntos de dados. Apesar da urgente necessidade de técnicas computacionais melhores, essa área ainda está relativamente atrasada em relação aos avanços realizados nas técnicas de obtenção de dados (LESK, 2007). Embora existam alguns repositórios úteis com ferramentas computacionais disponíveis para o pesquisador (e.g. Bioconductor (Fred Hutchinson Cancer Research Center, 2015)), e outras ferramentas estejam disponíveis a partir de sites dos próprios autores (LAKSHMANAN et al., 2012; GEVORGYAN et al., 2011; EFRONI et al., 2011; SHLOMI et al., 2008), muitas dessas técnicas não possuem interfaces de usuário intuitivas e exigem do pesquisador biólogo uma familiaridade com aspectos metacomputacionais que em geral ele não possui. Há uma necessidade grande de que as ferramentas computacionais para a biologia sejam desenvolvidas de forma multidisciplinar, em constante interação com os principais interessados, os investigadores da biologia (WOOLEY; LIN, 2005). É provável que em diversos trabalhos de pesquisa, devido às dificuldades associadas com o uso das ferramentas computacionais disponíveis, as ferramentas sejam selecionadas e utilizadas de maneira acrítica, em geral pela popularidade da ferramenta. Todavia, essa escolha é por vezes incorreta, produzindo resultados que são não apenas inúteis, mas errados. Apenas como exemplo de que ferramentas computacionais são altamente necessárias na pesquisa biológica atual, um

dos artigos de biologia mais citados na década precedente (LESK, 2007) foi o artigo de apresentação de uma técnica computacional para busca de informações biológicas por algoritmos de busca de similaridade e alinhamento entre estruturas amino-acídicas e ácido-nucleicas, o BLAST (*Basic Local Alignment Search Tool*) (ALTSCHUL et al., 1990).

Ainda um outro ponto importante, também associado à obtenção de grandes conjuntos de dados, criação de bases de dados e técnicas de mineração de dados é o aparecimento das redes genéticas e metabólicas. Esses dois tipos de redes, entretanto, são obtidos por processos bastante distintos. Enquanto uma rede metabólica é essencialmente o agrupamento de reações bioquímicas identificadas em experimentos de bancada isolados, as redes de regulação e sinalização são inicialmente estimadas a partir de dados de micromatrizes e então validadas com experimentos *in vitro*.

1.3 BACTÉRIA E METABOLISMO

As técnicas computacionais desenvolvidas neste trabalho assumem um comportamento celular para o *Mycobacterium tuberculosis* que pode ser representado pelo diagrama na Fig. 2. Esse processo, desde o sensoreamento celular das condições ambientais até a alterações no metabolismo, permite que as células se adaptem a situações de estresse causadas por fatores ambientais como depleção de nutrientes e presença de antibióticos.

Conforme boletim da Organização Mundial da Saúde, aproximadamente um bilhão de dólares anuais são gastos para tratamento da tuberculose em 22 dos países com maior incidência da doença (FLOYD; PANTOJA; DYE, 2007). Comparativamente, um quarto desse valor (aproximadamente 250 milhões de dólares) são gastos por ano em investimentos de pesquisa sobre a doença, conforme relatório do TAG *Treatment Action Group* (FRICK; JIMÉNEZ-LEVI, 2013), um dos parceiros mundiais do Grupo *Stop TB Partnership*. Certamente os custos com tratamento e hospitalização poderiam ser significativamente reduzidos se possuíssemos um maior entendimento da fisiologia da bactéria causadora da doença e da sua capacidade de adaptação e mutação. Nesse ensejo, se modelos preditivos mais compreensivos estivessem disponíveis, esses poderiam ser usados para estudar o efeito de novas drogas sobre as células doentes ou sobre as bactérias agentes. Esses modelos biológicos necessariamente serão construídos de forma iterativa. Assim, dados experimentais treinam modelos e validam suas predições, e por sua vez

estas predições podem ser usadas para definir novos experimentos úteis para a próxima iteração do processo de modelagem (PALSSON, 2000).

A Figura 2 mostra qual aspecto do metabolismo é necessário estudar para que novos compostos antibióticos possam ser desenvolvidos. Neste trabalho assume-se que a bactéria, ao sofrer exposição por um agente estressante, modifica o seu metabolismo de forma a aumentar as suas chances de sobrevivência. Assume-se ainda que essas alterações metabólicas ocorrem a nível de caminhos metabólicos, isto é, que o organismo diminui a atividade metabólica em caminhos que não são necessários à sobrevivência, e aumenta os níveis de atividade em caminhos metabólicos que são mais úteis à manutenção da vida do organismo. Estes últimos denominamos caminhos metabólicos de sobrevivência, ou apenas caminhos de sobrevivência.

É conhecido o fato que o *Mycobacterium tuberculosis* possui alta plasticidade metabólica, e que de forma recorrente desenvolve resistência aos antibióticos de linha de frente utilizados no tratamento da tuberculose. O surgimento, em várias regiões do mundo onde a tuberculose ainda causa número alarmante de mortes (YOUNG et al., 2008), de cepas MDR (*multi drug resistant*) e XDR (*extensively drug resistant*) (TB Alliance, 2000; STEENWINKEL et al., 2010) apenas aumenta a necessidade de compreender melhor os mecanismos genéticos e metabólicos utilizados pelo organismo para evadir a ação antibiótica dos compostos e, eventualmente, desenvolver resistência. São necessários modelos computacionais que permitam entender a fisiologia do organismo não apenas em situações regulares de homeostase, mas principalmente que permitam elucidar o metabolismo de sobrevivência. Precisa-se de modelos que permitam compreender e predizer o mecanismo de ação de novos compostoso antibióticos candidatos e os possíveis mecanismos de resistência que são utilizados pela bactéria.

Nos próximos capítulos serão apresentados e discutidos alguns procedimentos computacionais que pretendem auxiliar no estudo do metabolismo. Um dos objetivos deste trabalho é o desenvolvimento e a discussão de ferramentas computacionais que possam ser aplicadas pelo investigador-biólogo no contexto da pesquisa de antibióticos. É importante que essas ferramentas sejam intuitivas e que os resultados obtidos com essas ferramentas sejam de fácil interpretação biológica.

Veremos que modelos mais detalhados e que sejam, ao mesmo tempo, mais específicos para o organismo produzem benefícios de predição. Esse aumento na qualidade do modelo de predição ficará aparente quando apresentarmos a nova função objetivo para ser usada na técnica de balanço de fluxos quando os organismos são expostos a compostos

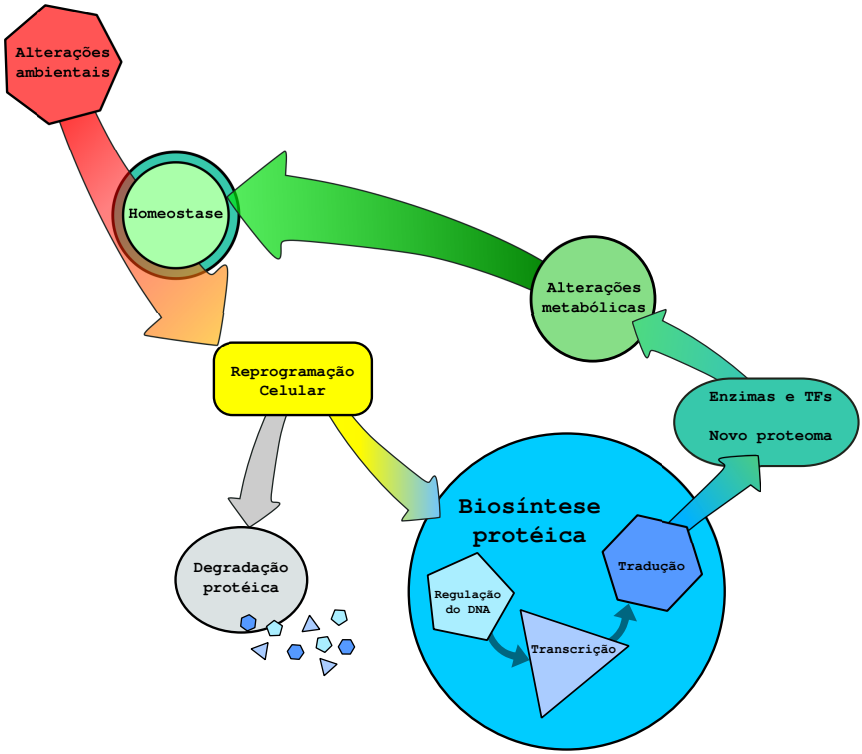

Figura 2 – A homeostase do organismo é perturbada pela exposição a situações de estresse nutricional, hipoxia/anoxia ou presença de antibióticos. Esses fatores ambientais iniciam um processo de reprogramação celular, onde proteínas que não são essenciais para a nova situação são degradadas e suas partes reutilizadas como bloco para síntese de novas proteínas. Um novo conjunto de proteínas é sintetizado para completar o proteoma necessário para lidar com a nova situação. É feita a leitura do DNA, transcrição de mRNAs e finalmente tradução dos RNAs mensageiros em proteínas funcionais. Essas novas proteínas, sejam elas fatores de transcrição ou de função enzimática, provocam as alterações necessárias no metabolismo para atingir novamente um grau de homeostase (em fenótipos de resistência) ou apenas sobrevivência temporária. *Fonte: Elaborada pelo autor.*

tóxicos.

Apresentaremos uma técnica que pode ser utilizadas na classificação de compostos antibióticos segundo o seu mecanismo de ação, e outro modelo matemático de regressão que permite identificar caminhos metabólicos utilizados pelo organismo no seu metabolismo de sobrevivência. Foi nosso intuito com este trabalho propor ferramentas que produzam pistas biológicas para o desenvolvimento de novas drogas sinergísticas para controle da tuberculose. Espera-se que, com o aprimoramento dos métodos apresentados, seja possível predizer quais caminhos metabólicos são mais prováveis de serem usados pela bactéria quando é exposta a um composto ou estresse biológico específico.

Assumindo que essas predições indiquem alvos que possam ser biologicamente manipulados, o desenvolvimento de uma nova droga de ação complementar permitirá bloquear estes caminhos de sobrevivência e diminuir o tempo de sobrevida das colônias bacteriais. Como o desenvolvimento de cepas resistentes está diretamente associado com a duração do tratamento (CASTELNUOVO, 2010), espera-se que a redução desse tempo leve ao processo de depleção completa das colônias antes que essas possam desenvolver resistência, de forma que os modelos e técnicas apresentados sejam a alavanca necessária na busca por novos caminho profiláticos para a TB.

2 EXPRESSÃO GÊNICA

2.1 ANÁLISE TRANSCRIPTÔMICA

Qualquer tentativa de utilizar dados experimentais de transcriptômica para a estimação de modelos biológicos esbarra na pequena quantidade de amostras produzida em cada experimento de micromatrizes de DNA. Os dois principais motivos para esse reduzido número de amostras são o alto custo por experimento e a complexidade do procedimento experimental. Na grande maioria dos experimentos, o número projetado de repetições para cada condição experimental, principalmente em estudos longitudinais, raramente ultrapassa 5 a 7 amostras. Entretanto, com a constante redução no custo da realização do procedimento de micromatrizes de DNA este número gradativamente tende a aumentar. Combinado com o fato que cada amostra possui alta dimensionalidade (alguns milhares de medições no caso de pequenos organismos procariotas), é importante o estudo de técnicas de avaliação da qualidade das estimativas produzidas com tais dados, e de subsequentes técnicas de geração de dados sintéticos usando as variáveis estimadas. Em estudos com um número n de amostras pequeno (i.e. poucas amostras) e cujos dados são de dimensão p muito alta (problemas p-grande, n-pequeno), ainda que seja essencial o uso de estimadores não assintóticos, frequentemente são usados estimadores de máxima verossimilhança de altíssima variância, que essencialmente não conseguem descrever o processo aleatório subjacente gerador dos dados.

Em muitas aplicações de análise de transcriptômica é necessário obter uma estimativa da matriz de covariância dos dados, ou da inversa dessa matriz. Por exemplo, assumindo que vetores de medições de expressão genética possam ser modelados como distribuições gaussianas multivariadas (BALDI; LONG, 2001; LEE, 2004):

$$\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S}) \tag{2.1}$$

Como a estimação do vetor de média em muitos casos não é crítica ou difícil, o problema da geração de dados de expressão genética sintéticos se resume ao problema de uma estimação satisfatória da matriz de covariância do processo. Note-se, que os milhares de elementos do vetor de expressão não podem, *a priori*, ser assumidos independentes ou descorrelacionados. Isso porque a expressão gênica de um gene pode estar fortemente associada à expressão de outro(s), devido, en-

tre outros motivos, à presença de conexões de regulação genética e à expressão conjunta de genes associados em *regulons*, Dessa forma, é essencial estimar uma matriz de covariância completa, com dimensão de centenas de milhares de elementos, utilizando poucas amostras, com uma razão p/n em geral muito maior que 10. Todavia, em situações práticas, comumente é realizada uma seleção de descritores prévia com um algoritmo de agrupamento ou outro critério de seleção por redução de redundância (BOSHOFF et al., 2004). Dessa forma, em problemas práticos essa relação não ultrapassa o limite $p/n = 10$. Por esse motivo, nas simulações que serão apresentadas utilizamos uma razão máxima igual a 10.

Embora o problema soe essencialmente sem solução, respostas são necessárias, e tentativas de uso de poucas amostras de dados em procedimentos de estimação tornam a aparecer (HOFFBECK; LANDGREBE, 1996; TADJUDIN; LANDGREBE, 1999) com o objetivo de aprimorar os resultados cada vez mais.

Nesta seção vamos revisar e avaliar a técnica de estimação por encolhimento (*shrinkage*). Essa técnica é importante por exemplo, para estimação da matriz de covariância de amostras com distribuição gaussiana multivariada. Uma aplicação é a estimação da matriz de covariância ou correlações de dados de expressão transcriptômica, os quais são, essencialmente, procedimentos de estimação *p-grande, n-pequeno*.

O método de *shrinkage* não é novo, e embora possa ser aplicados na estimação de qualquer variável associada a um conjunto de dados aleatórios, grande parte dos estudos foca em como aprimorar a estimação da matriz de covariância de uma variável aleatória normal multivariável.

O foco aqui será mostrar que a técnica de *shrinkage* é especialmente útil no caso de problemas com poucas amostras de experimentos de micromatrizes de DNA de alta dimensionalidade.

No caso de estimação da matriz de covariância, é essencial que o estimador produza uma matriz positiva definida e bem condicionada. Todavia, no caso de dados de transcriptômica, a estimação apresenta dificuldades oriundas do aspecto p-grande-n-pequeno do problema e, portanto, não é possível garantir nem uma nem outra característica. Surpreendentemente, mesmo com tais dificuldades, em muitos problemas práticos utiliza-se simplesmente o estimador empírico não-polarizado:

$$\hat{\boldsymbol{S}} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{e}_k - \bar{\boldsymbol{e}})(\boldsymbol{e}_k - \bar{\boldsymbol{e}})^T \tag{2.2}$$

em que $\boldsymbol{e}_k$ é uma observação multivariável do conjunto de n amostras e $\bar{\boldsymbol{e}}$ é o vetor da média amostral do conjunto de dados. Cada observação $\boldsymbol{e}_k$ pode representar, por exemplo, as medições obtidas com um experimento de micromatriz de DNA em uma determinada condição experimental. Embora de uso corrente, mesmo em problemas de $p \gg n$, uma estimativa obtida com esse estimador não pode ser considerada como uma boa aproximação da matriz de covariância verdadeira da população. Essa aproximação só será válida na medida que $n \approx p$. Há vários problemas na estimação da matriz de covariância usando o estimador $\hat{\boldsymbol{S}}$ segundo a expressão da Eq. (2.2) quando o número de amostras n é menor que a dimensão p da variável aleatória:

- posto incompleto, visto que $p - n$ autovalores da matriz estimada serão sempre nulos. Por exemplo, ao estimar a matriz de covariância de vetores com dimensão $p = 2000$ utilizando um número de amostras de micromatrizes $n = 100$, dos 2000 autovalores da matriz de covariância (ainda que muitos deles sejam muito pequenos), apenas 100 serão não-nulos. Um outro exemplo com menores dimensões está ilustrado na Fig. 3 para diferentes razões p/n;

- a matriz estimada não é positiva definida, visto que vários autovalores serão nulos;

- a matriz estimada apresenta mau condicionamento. Assumindo que $n < p$, quanto menor o número de amostras, maior a quantidade de autovalores nulos e maior a magnitude dos autovalores restantes. O condicionamento da matriz é degradado cada vez mais;

- não é possível inverter a matriz estimada.

Ainda que estes quatro problemas estejam intimamente relacionados e a rigor sejam equivalentes, ainda assim acreditamos ser útil identificá-los separadamente. É possível que em diferentes situações um destes aspectos seja preponderante ou mais intuitivo no momento de avaliar a qualidade do estimador.

Observar a norma de Frobenius da diferença das matrizes estimada e verdadeira deixa claro a redução na qualidade da estimativa para valores de n cada vez menores. Para uma quantidade de amostras igual à dimensão do vetor (razão $p/n = 1$) a norma da diferença em uma realização é igual a 4.8. À medida que a razão p/n aumenta para

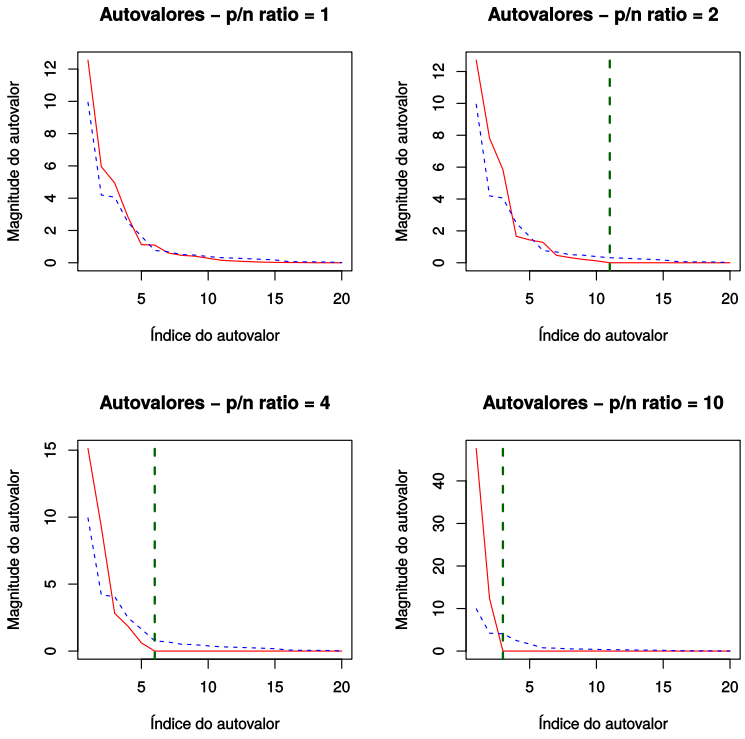

Figura 3 – Autovalores da matriz de covariância mostrados em ordem decrescente de magnitude ($p = 20$). Linha azul tracejada: matriz de covariância verdadeira. Linha vermelha: matriz de covariância ML. À medida que a relação $p/n > 1$ aumenta, e o número de amostras é menor que a dimensão do vetor, um número $p - n$ de autovalores da matriz estimada é igual a zero. Os autovalores positivos por sua vez apresentam magnitude cada vez maior, distanciando-se dos verdadeiros autovalores. Todos autovalores estimados a partir da linha tracejada vertical verde são nulos. *Fonte: Elaborada pelo autor.*

2, 4 e 10, a norma da diferença na mesma realização também aumenta para respectivamente 8.01, 9.9 e 41.2.

Visto que em situações de $p > n$ e $p \gg n$ o estimador $\hat{\boldsymbol{S}}$ não é satisfatório, são necessárias alternativas de estimação para a matriz de covariância. É importante que esses novos estimadores possuam as características desejáveis de serem positivos-definidos e bem condicionados. Como essas duas características estão relacionadas ao comportamento dos autovalores da matriz estimada, é possível utilizar, além do erro, o gráfico dos autovalores ordenados para avaliação da qualidade do estimador. Para que o estimador seja positivo-definido, todos os seus autovalores devem ser maiores que zero. Além disso, para que a matriz estimada apresente bom condicionamento, a razão entre o maior autovalor e o menor autovalor não deve tender para valores muito grandes quando n diminui. É possível ver na Fig. 3 que o estimador $\hat{\boldsymbol{S}}$ não apresenta essas características quando o número de amostras é muito pequeno.

Há diversos métodos que podem ser utilizados para melhorar a estimação dos autovalores e garantir uma matriz positiva definida e bem condicionada. Uma primeira alternativa bastante simples é ajustar a magnitude de todos os autovalores estimados para que eles sejam maiores que um determinado valor de limiar selecionado. Embora essa alternativa garanta uma estimativa positiva definida, em geral ela não corrige o condicionamento da matriz.

Uma segunda opção pode ser o uso de técnicas de *bootstrap* ou MCMC. Porém, tais técnicas, embora apresentando melhoria em relação ao estimador empírico não-polarizado, são computacionalmente exigentes (principalmente para valores de p muito altos como é o caso de dados de transcriptômica) além de necessitarem de suposições sobre a distribuição dos dados subjacente.

Uma terceira possibilidade de melhoria em relação ao estimador empírico é o uso de *shrinking* (SCHÄFER; STRIMMER, 2005), ou encolhimento, que além de produzir uma estimação da matriz de covariância positiva definida e com um melhor condicionamento que $\hat{\boldsymbol{S}}$, é um procedimento com menores exigências computacionais. A técnica de encolhimento é conhecida também como estimação polarizada, pois busca-se diminuir a variância da estimativa permitindo que ela seja polarizada, o que em casos de poucas amostras de alta dimensionalidade é um aspecto bastante atraente da técnica, visto que outros estimadores teriam variância alta demais para que fossem realmente úteis.

2.2 ESTIMAÇÃO POR ENCOLHIMENTO

Basicamente, o método de encolhimento (*shrinkage*) produz uma estimativa a partir da mistura ponderada de dois estimadores (SCHÄFER; STRIMMER, 2005): um estimador *sem restrições* e um outro estimador, chamado estimador alvo, que possui restrições que diminuem a quantidade de parâmetros a serem estimados, porém acrescentando polarização à estimativa. O estimador sem restrições em geral é um estimador assintótico de máxima verossimilhança, o qual é ótimo se o número de amostras é suficientemente grande. Um exemplo que permite compreender melhor os componentes da mistura é dado a seguir. Um exemplo de estimação sem restrições é quando deseja-se estimar uma matriz de covariância de estrutura desconhecida com tamanho $p \times p$, em que é necessário estimar todos os $p^2/2 + p/2$ parâmetros (variâncias e covariâncias). Todavia, podemos assumir que a estrutura da matriz apresente todas as covariâncias nulas, e essa restrição reduz o número de parâmetros a estimar para apenas p, (i.e. os elementos de variância da diagonal). Claramente, para um mesmo número de amostras n, a variância da estimativa será menor, porém às custas de uma polarização em todos os parâmetros de covariância. Outra possibilidade poderia ser assumir, para o estimador-alvo, que todas as suas variâncias sejam iguais (uma única variância comum ao longo da diagonal da matriz), e que todas as covariâncias também o sejam (uma covariância comum para todos os elementos fora da diagonal). Nesse caso, o número de parâmetros a estimar teria sido reduzido a apenas dois: a variância comum e a covariância comum. Entretanto, mais uma vez, as estimativas serão polarizadas.

A ideia do encolhimento é não usar nem a estimativa de alta variância sem restrições, nem a estimativa polarizada com restrições, mas uma média ponderada desses dois estimadores, em que o peso de cada componente é dado pelo índice de encolhimento λ. Usando a notação $\boldsymbol{U}$ para o estimador sem restrições e $\boldsymbol{T}$ para o estimador-alvo com restrições, podemos escrever o estimador por encolhimento $\boldsymbol{S}^*$ como:

$$\boldsymbol{S}^* = \lambda \boldsymbol{T} + (1 - \lambda)\boldsymbol{U} \quad , \qquad 0 \leq \lambda \leq 1 \tag{2.3}$$

A estimativa regularizada $\boldsymbol{S}^*$ tem características intermediárias, e portanto melhores em relação tanto a $\boldsymbol{T}$ quanto a $\boldsymbol{S}$. Maiores valores do parâmetro λ permitem que mais peso seja dado ao estimador-alvo, i.e. às estimativas polarizadas, o que é útil nos casos de um número pequeno de amostras que produzem uma estimativa assintótica de alta

variância. Valores menores para λ dão mais peso para o estimador sem restrições, em geral quando um maior número de amostras está disponível.

Pela Eq. (2.3), assumindo que $\boldsymbol{U}$ seja o estimador de máxima verossimilhança $\hat{\boldsymbol{S}}$, vê-se que há dois graus de liberdade no cálculo da estimativa por encolhimento:

- seleção do estimador-alvo $\boldsymbol{T}$
- determinação do nível de encolhimento λ

Dos dois itens acima, a determinação de valor ótimo para o parâmetro de encolhimento λ é o mais crítico. Embora o valor de λ possa ser definido simplesmente escolhendo-se um valor pré-fixado no intervalo $[0, 1]$, ou mesmo como uma função simples do número de amostras, um valor ótimo λ^* pode ser mais adequadamente encontrado através de um procedimento de minimização, validação cruzada, ou pode ainda ser determinado analiticamente, conforme mostrado a seguir.

Segundo (SCHÄFER; STRIMMER, 2005), a expressão analítica que garante o menor erro quadrático médio para o estimador por encolhimento $\boldsymbol{S}^*$ é devida a um resultado apresentado por Ledoit e Wolf em 2003 (LEDOIT; WOLF, 2003), e dada por:

$$\lambda^* = \frac{\sum_{i=1}^{p} \operatorname{var}(s_i) - \operatorname{cov}(t_i, s_i)}{\sum_{i=1}^{p} E\{(t_i - s_i)^2\}} \tag{2.4}$$

em que o estimador sem restrições $\boldsymbol{U}$ foi substituído pela estimador empírico não-polarizado da Eq. (2.2) e a notação $E\{\cdot\}$ indica o valor esperado da variável aleatória indicada no argumento. Essa expressão é uma versão simplificada da versão dada em (LEDOIT; WOLF, 2003) para o caso em que $\boldsymbol{U}$ é um estimador não-polarizado.

Entretanto, para encontrar o valor ótimo do parâmetro de encolhimento ainda é necessário definir o estimador-alvo, i.e. o estimador com restrições polarizado $\boldsymbol{T}$. Em geral, a escolha desse estimador não é tão crítica quanto a seleção do parâmetro ótimo λ^*, e vários alvos podem ser utilizados que garantam uma melhoria da estimação da matriz. Essa escolha não é crítica porque a própria expressão para λ^* já incorpora mecanismos que evitam uma degradação significativa no caso de alvos mal especificados.

Segundo (SCHÄFER; STRIMMER, 2005), uma característica importante da Eq. (2.4) é que quanto menor a variância da estimativa sem restrições $\boldsymbol{U}$ (i.e. maior número de amostras n disponível usadas na estimação), menor o valor ótimo λ^*. De outra forma, o valor do

parâmetro também depende do erro quadrático médio entre os estimadores com e sem restrições, diminuindo o valor de λ^* com o aumento do MSE, evitando assim que a estimativa $\boldsymbol{S}^*$ sofra degradação por um estimador-alvo mal especificado. Finalmente, como as variâncias, covariâncias e valor esperado indicados na Eq. (2.4) não estão disponíveis, (LEDOIT; WOLF, 2003) sugere utilizar as suas estimativas amostrais não-polarizadas.

A determinação do estimador-alvo, como mencionado, não é uma questão crítica, e melhoras significativas no erro de estimação e na variância das estimativas podem ser obtidas com alvos $\boldsymbol{T}$ diversos. Todavia, segundo (SCHÄFER; STRIMMER, 2005), no caso de estimação de matrizes de covariância para dados de expressão genética, o alvo diagonal com variâncias desiguais é uma escolha em geral apropriada. Independentemente da escolha do estimador-alvo, sempre há redução no erro da norma de Frobenius, a versão do MSE para matrizes de covariância. É fato que, para a escolha de um estimador-alvo que não corresponda minimamente à estrutura subjacente que imaginamos para a matriz de covariância da população, o estimador $\boldsymbol{S}^*$ será pouco eficiente, e consequentemente a redução no MSE será pequena. Visto que em tal situação o parâmetro de encolhimento λ^* terá um valor muito pequeno, o resultado será que $\boldsymbol{S}^* \approx \hat{\boldsymbol{S}}$.

Dentre alguns dos possíveis alvos mais usados destacam-se a matriz identidade e seus múltiplos escalares (HOFFBECK; LANDGREBE, 1996), além do modelo já mencionado com variância e covariância comuns. Duas propriedades desses estimadores são: (1) a sua baixíssima dimensionalidade (i.e. 1 ou 2 parâmetros a estimar), o que permite uma estimação eficiente mesmo com n muito pequeno, e (2) o alcance do encolhimento, que atua sobre todos os elementos da matriz, tanto variâncias quanto covariâncias. Todavia, ainda segundo (SCHÄFER; STRIMMER, 2005), um estimador-alvo do tipo "diagonal com variâncias distintas"é uma opção mais adequada. A complexidade computacional, ainda que significativamente maior dependendo do valor de p, não é tão alta quanto a de estimadores amostrais, e permite que o encolhimento seja aplicado apenas às covariâncias, sem encolher as variâncias na diagonal. Essa capacidade de atuar diferentemente sobre dois grupos de parâmetros traz mais flexibilidade à técnica. Um benefício adicional é que esse estimador-alvo sempre produz uma estimativa positiva definida, o que não é garantido, por exemplo, para o estimador biparamétrico de variância e covariância comuns.

Para demonstrar as melhorias significativas derivadas do uso do encolhimento na estimação de matrizes de covariância para dados de

expressão genética, e seguindo as indicações de (SCHÄFER; STRIMMER, 2005), na seção seguinte apresentamos alguns exemplos que utilizam como estimador-alvo $\boldsymbol{T}$ a matriz "diagonal com variâncias distintas", em que todos os elementos de covariância são nulos.

2.3 RESULTADOS

O uso de *shrinkage* (i.e. encolhimento) na estimação de matrizes de covariância tem por objetivo produzir um estimador que compense a tendência de afastamento dos autovalores observada na Fig. 3. O encolhimento minimiza grandemente os problemas identificados na seção anterior com o estimador empírico, como posto incompleto, autovalores nulos e mau condicionamento.

Com o encolhimento, mesmo que a estimação seja feita com menos amostras que a dimensão p, a energia é espalhada por todos os autovalores, de forma que não há valores nulos, como no caso da estimativa amostral. Além disso, há uma significativa redução na diferença entre as normas de Frobenius da matriz verdadeira $\boldsymbol{S}$ e da matriz estimada $\hat{\boldsymbol{S}}_{ML}$, e uma melhora no número de condicionamento como veremos nos exemplos a seguir.

Na Fig. 4 temos uma simulação comparando as magnitudes ordenadas dos autovalores dos estimadores ML (máxima verossimilhança) e SH (*shrinkage*) e os valores da matriz de covariância verdadeira da população. Para esse exemplo temos $p = 20$ e quatro razões p/n que mostram claramente a degradação causada na estimação à medida que p/n aumenta.

O comportamento dos autovalores estimados com a redução de n pode ser descrito como um *afastamento* em relação aos autovalores verdadeiros, e que é dependente da magnitude do autovalor. Autovalores maiores (mais à esquerda na figura) aumentam de magnitude, enquanto autovalores menores diminuem a sua magnitude (se $n < p$ vários serão nulos). Observa-se um ponto que divide o tipo de afastamento. Por exemplo, no primeiro exemplo da Fig. 3, para uma razão $p/n = 2$, o afastamento positivo (aumento de magnitude) ocorre apenas para os cinco maiores autovalores, enquanto os demais apresentam afastamento negativo (diminuem de magnitude). O comportamento exato de uma realização depende, claramente, dos dados. As consequências mais importantes decorrentes do uso do estimador por encolhimento que podem ser diretamente observadas com a simulação da Fig. 4 são uma significativa redução no erro quadrático (norma de Frobenius da diferença

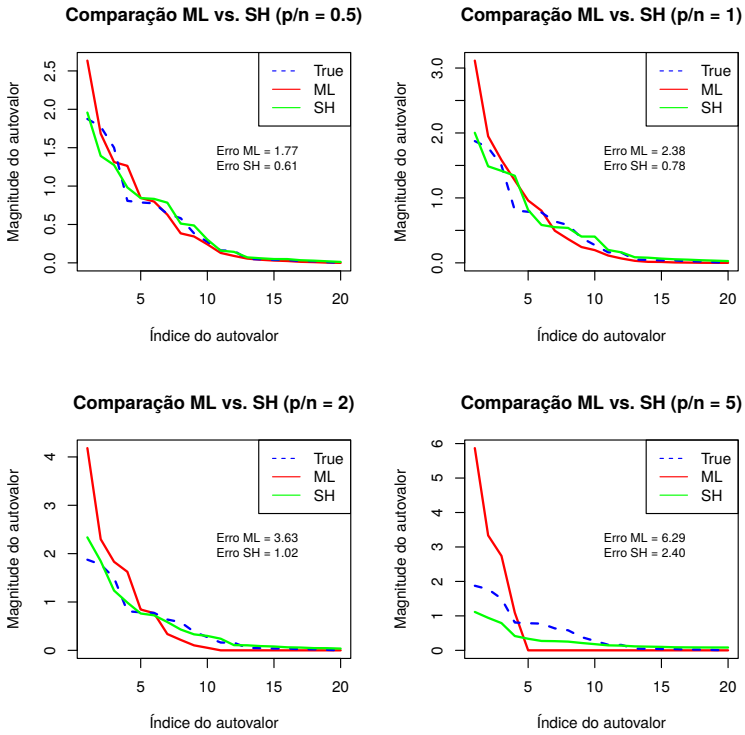

Figura 4 – Autovalores da matriz de covariância em ordem decrescente de magnitude. Simulação dos estimadores empírico e com encolhimento para cada uma das razões p/n: 0.5, 1, 2, 5, para $p = 20$. Os autovalores verdadeiros estão mostrados em azul. Observa-se claramente nesta realização que o estimador por encolhimento diminui o afastamento dos autovalores estimados em relação aos verdadeiros, melhora o condicionamento da matriz estimada e apresenta significativa redução no erro de estimação (norma de Frobenius da diferença da matriz estimada e verdadeira). *Fonte: Elaborada pelo autor.*

entre matriz estimada e matriz verdadeira), redução no número de autovalores nulos para razões $p/n > 1$ (garantindo posto completo para a matriz estimada), e uma diminuição no erro de estimação dos autovalores de maior magnitude, produzindo um melhor condicionamento da matriz estimada.

É importante avaliar a qualidade e características do estimador em um conjunto de várias realizações, de forma que nos próximos exemplos observaremos o comportamento médio e variância dos estimadores ao longo de diversas realizações.

Para o estimador empírico não-polarizado $\hat{\boldsymbol{S}}$, novamente com $p = 20$, foram simuladas $R = 1600$ realizações do processo de estimação para as quatro diferentes razões p/n (1, 2, 4 e 10), e calculada a norma de Frobenius (norma-F) da diferença entre as matrizes verdadeira e estimada. Os histogramas resultantes estão mostrados na Fig. 5. Nessa figura pode-se observar que a distribuição da diferença entre as normas segue uma distribuição assimétrica.

Com o aumento da razão p/n (redução no número de amostras), há um aumento da distância entre a média da distribuição e o valor zero, indicando uma estimação menos precisa, além de um consoante aumento na variância da variável aleatória. A assimetria da distribuição é resultante do maior aumento de magnitude dos autovalores maiores, e da convergência dos $p - n$ autovalores menores para zero.

Um estimador com encolhimento (*shrinkage*), ainda que sofra degradação com a diminuição do número de amostras, produz estimativas muito mais precisas, com menor variância e com uma melhor distribuição da energia por todos os autovalores. Esses resultados são mostrados na Fig. 4, onde observa-se uma redução no afastamento dos autovalores em relação às magnitudes da matriz verdadeira e na Fig. 6, a qual mostra os histogramas a norma-F da diferença para a estimação com encolhimento.

Um outro aspecto importante da estimação por encolhimento diz respeito ao condicionamento da matriz $\boldsymbol{S}^*$ estimada. Para o estimador empírico não-polarizado da Eq. (2.2) não é possível analisar o condicionamento para casos $n < p$, visto que a matriz estimada é singular. Já para o estimador por encolhimento, o bom condicionamento da estimativa é patente. Na simulação da Fig. 7, a matriz por encolhimento foi estimada para $R = 1600$ realizações, com $p = 100$. Os valores singulares da estimativa foram determinados pela decomposição SVD e o número de condicionamento, i.e. a razão entre o valor singular máximo e o valor singular mínimo, calculados para cada realização. Os histogramas dos números de condicionamento obtidos mostram que mesmo

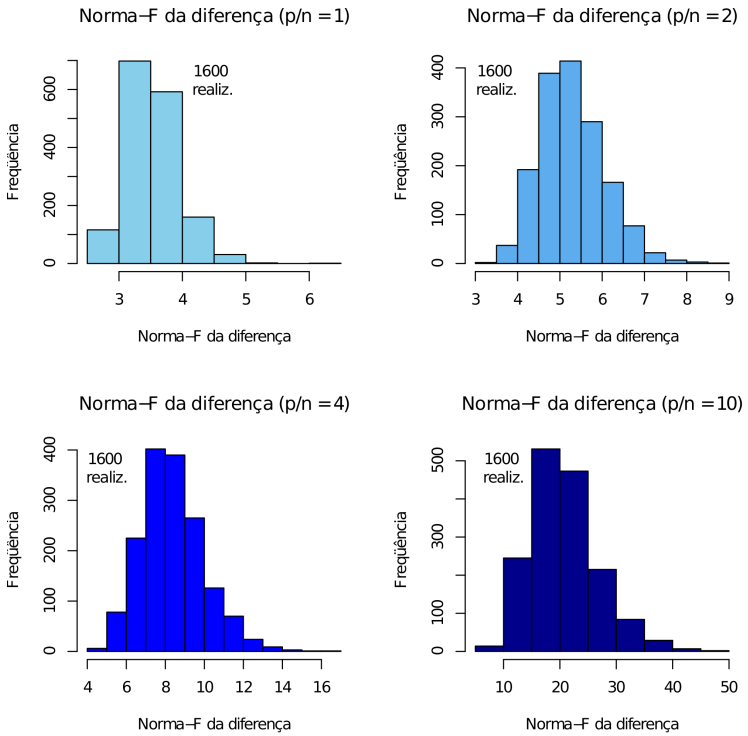

Figura 5 – Histogramas da norma de Frobenius (norma-F) da diferença entre as matrizes verdadeira $\boldsymbol{S}$ e estimada $\hat{\boldsymbol{S}}$. Foram realizadas $R = 1600$ simulações do estimador para cada uma das razões p/n: 1, 2, 4, 10, em que $p = 20$. Observa-se claramente a assimetria da distribuição. À medida que o número de amostras diminui, a diferença entre as duas aumenta. *Fonte: Elaborada pelo autor.*

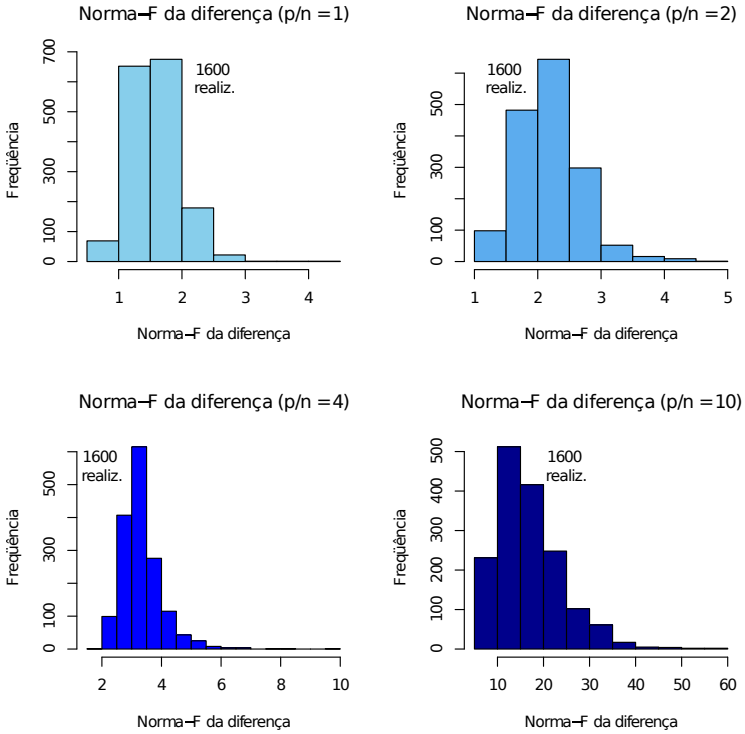

Figura 6 – Histogramas da norma de Frobenius (norma-F) da diferença entre as matrizes verdadeira $\boldsymbol{S}$ e estimada $\boldsymbol{S}^*$. Foram realizadas $R = 1600$ simulações do estimador para cada uma das razões p/n: 1, 2, 4, 10, em que $p = 20$. A estimação com encolhimento produz erros menores que o estimador empírico para todas as quantidades de amostras simuladas. *Fonte: Elaborada pelo autor.*

Tabela 1 – Médias e desvios-padrão da norma da diferença para o estimador empírico e por encolhimento para $p = 20$.

p	p/n	μ_{ML}	μ_{SH}	sd_{ML}	sd_{SH}
20	1	2.36	1.33	0.26	0.27
20	2	3.53	1.86	0.54	0.38
20	4	5.61	2.88	1.09	0.59
20	10	14.09	14.39	4.30	5.60

para casos com número de amostras muito menor que a dimensão do vetor de dados observados, o condicionamento da matriz de covariância estimada é garantido. Observa-se na figura que para $p = 100$, razões p/n maiores tendem a aumentar o valor do parâmetro de encolhimento ótimo λ^*, aumentando a taxa de encolhimento (i.e. afastamento dos autovalores) e, consequentemente melhorando o condicionamento.

É interessante notar que o comportamento do condicionamento da matriz é dependente da dimensão p do problema. Para o caso com $p = 20$, as simulações para as mesmas razões p/n anteriores resultam em degradação do condicionamento. Essa tendência é resultante do número muito pequeno de amostras usado na estimação, mesmo para o estimador por encolhimento. Na Fig. 8, para $p/n = 10$ o condicionamento é perdido, como no caso do estimador empírico, visto que apenas duas obervações estão disponíveis para estimar uma matriz 20×20. Claramente, em casos extremos, mesmo a técnica por encolhimento apresenta seus limites.

Uma última observação interessante diz respeito à dependência da variância das estimativas em relação à dimensão do problema. Quanto maior a dimensionalidade p dos dados observados, menor a variância das estimativas, tanto para o estimador empírico quanto para o estimador por encolhimento. Na Tabela 1, apresentamos para os dois estimadores a média e a variância da norma-F da diferença para valor de $p = 20$. Na Tabela 2, estão os mesmos resultados de simulação agora para $p = 200$. Diferentemente da simulação anterior, o número de amostras n para a maior razão são 20 observações. Vê-se que, nesse caso, o estimador por encolhimento não é degradado como no caso de $p = 20$, em que apenas duas amostras estavam disponíveis para estimação de 400 parâmetros.

Finalmente, uma comparação das médias e desvios padrão para a norma-F da diferença entre a matriz verdadeira e estimada, para uma

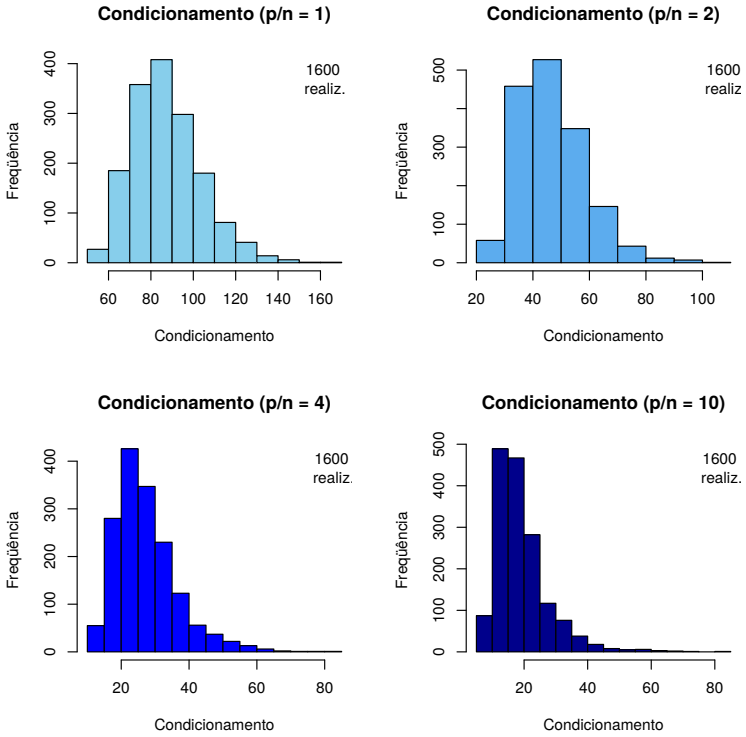

Figura 7 – Estimação por encolhimento. Histogramas do número de condicionamento (razão entre os valores singulares máximo e mínimo) da matriz estimada $\boldsymbol{S}^*$. Foram realizadas $R = 1600$ simulações do estimador para cada uma das razões p/n: 1, 2, 4, 10, em que $p = 100$. A estimação com encolhimento produz estimativas com razoável condicionamento mesmo para conjuntos de amostra com n menor que a dimensão p do problema. *Fonte: Elaborada pelo autor.*

Tabela 2 – Médias e desvios-padrão da norma da diferença para o estimador empírico e por encolhimento para $p = 200$.

p	p/n	μ_{ML}	μ_{SH}	sd_{ML}	sd_{SH}
200	1	11.56	1.32	0.13	0.10
200	2	16.44	1.86	0.24	0.14
200	4	23.48	2.61	0.46	0.20
200	10	38.28	4.24	1.15	0.29

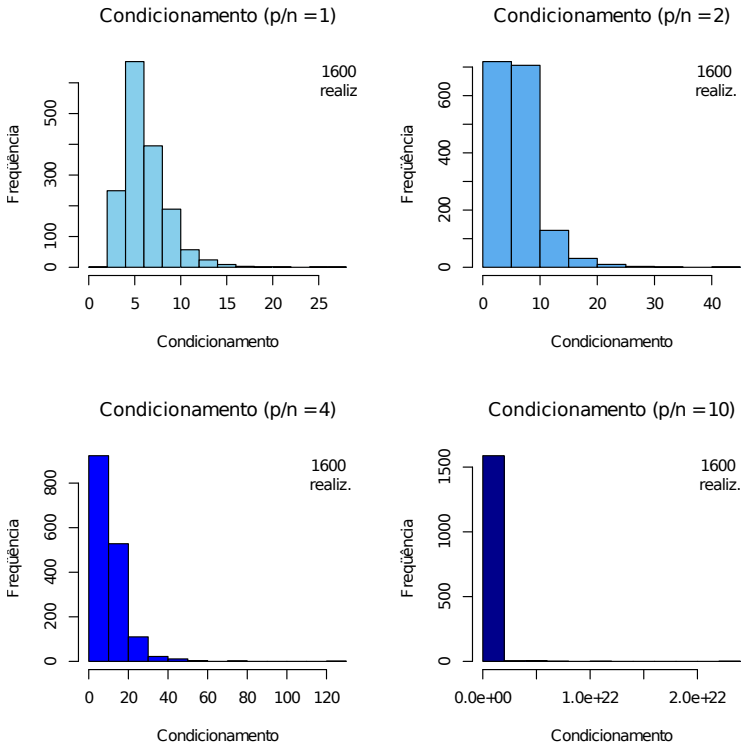

Figura 8 – Estimação por encolhimento. Histogramas do número de condicionamento (razão entre os valores singulares máximo e mínimo) da matriz estimada $\boldsymbol{S}^*$. Foram realizadas $R = 1600$ simulações do estimador para cada uma das razões p/n: 1, 2, 4, 10, em que $p = 100$. A estimação com encolhimento produz estimativas com razoável condicionamento mesmo para conjuntos de amostra com n menor que a dimensão p do problema. *Fonte: Elaborada pelo autor.*

Tabela 3 – Médias e desvios-padrão da norma da diferença para o estimador empírico e por encolhimento para diferentes valores de p e razão $p/n = 4$.

p	p/n	μ_{ML}	μ_{SH}	sd_{ML}	sd_{SH}
20	4	5.61	2.89	1.09	0.59
100	4	17.29	2.65	0.64	0.27
200	4	23.48	2.60	0.46	0.20
500	4	34.06	2.80	0.27	0.14
1000	4	50.34	2.87	0.19	0.10

mesma razão p/n e diferentes valores de p é apresentada na Tabela 3 para os dois estimadores. Observamos nesses dados que para ambos os estimadores a variância do erro (i.e. norma-F da diferença entre matriz estimada e verdadeira) diminui com o aumento da dimensionalidade. A magnitude média do erro aumenta com a dimensão p para ambos estimadores. Todavia, é claro observar a vantagem do estimador por encolhimento em relação ao estimador empírico, pois aquele apresenta tanto menor valor médio para o erro quanto menor variância ao redor dessa média, quando comparado com o estimador empírico. Além disso, a variação do erro com a dimensão é muito menor que para o estimador empírico. Ao passo que para o estimador empírico um aumento de 50 vezes na dimensão resultou em um aumento de aproximadamente 10 vezes no erro quadrático médio, para o estimador por encolhimento, o mesmo aumento na dimensionalidade do problema não resultou em nenhum aumento significativo do erro. Dessa forma, é altamente recomendável o uso do estimador por encolhimento, visto que este produz estimativas muito mais precisas. Embora nessa tabela tenham sido apresentados dados apenas para razão $p/n = 4$, em geral as conclusões são semelhantes para outras razões menores ou maiores. Tais comparações entretanto podem ser feitas apenas dentro de certos limites, pois é possível que o estimador por encolhimento não apresente vantagens em situações extremas como uma razão $p/n = 10$ para $p = 20$, quando apenas duas observações estão disponíveis.

Podemos também observar o comportamento médio do condicionamento da matriz estimada para o caso do encolhimento. Não é possível estudar esse parâmetro no caso do estimador empírico para razões p/n maiores que a unidade. Na Tabela 4 apresentamos os resultados para o número de condicionamento de simulações do estimador por encolhimento para diferentes valores de p e diversas razões p/n.

Tabela 4 – Médias e variâncias do condicionamento do estimador por encolhimento para diferentes valores de p e diversas razões p/n.

p	p/n	μ_{SH}	sd_{SH}
20	1	18.22	7.93
20	2	14.76	9.19
20	4	22.61	18.47
20	10	∞	∞
100	1	58.43	10.17
100	2	38.46	8.08
100	4	24.99	6.75
100	10	19.62	7.26
200	1	154.74	14.53
200	2	81.77	10.20
200	4	45.00	7.25
200	10	24.25	5.72
500	1	204.59	11.86
500	2	134.79	10.12
500	4	82.95	8.03
500	10	42.85	5.87
1000	1	791.05	37.60
1000	2	452.33	27.54
1000	4	246.51	19.90
1000	10	109.48	12.84

Observa-se que a faixa dinâmica dos valores singulares aumenta para uma mesma razão p/n e maiores valores de p. Para uma dimensionalidade fixa, a tendência do encolhimento é comprimir a faixa dinâmica dos valores singulares, diminuindo o valor do condicionamento, melhorando as características da matriz para inversão, ainda que haja introdução de uma polarização nas estimativas. Entretanto, é importante ressaltar que o erro quadrático médio do estimador por encolhimento é significativamente menor que para o estimador empírico, e ainda apresenta bom condicionamento, permitindo que obtenhamos a inversa da matriz estimada sem sofrer com os problemas numéricos e singularidade do estimador empírico.

Para os casos mais extremos de $p = 20$, o comportamento do condicionamento é aparentemente inverso ao das simulações com p maior. Nesse caso, tanto a média quanto a variância do condicionamento pa-

recem aumentar com o aumento da dimensionalidade. É possível que, devido ao número muito reduzido de observações, o condicionamento da estimativa fique comprometido.

Para complementar os resultados do capítulo, comparamos a estimação empírica e por encolhimento dos coeficientes de correlação de um conjunto de 700 genes de proteínas com atividade enzimática utilizando apenas $n = 23$ amostras de um experimento de micromatriz de DNA com um grupo de controle ($n_c = 11$ para quatro instantes amostrais[1]) e outro de tratamento ($n_t = 12$ com três repetições para cada instante amostral) com exposição do Mtb ao composto mefloquina, um possível candidato a antibiótico. A razão p/n para este problema é mais alta que a apresentada nas simulações anteriores. Nas simulações foram estimadas tanto a matriz de covariância quanto a matriz de correlação dos dados.

Confirmando os resultados apresentados anteriormente neste capítulo, a qualidade da estimação da matriz de covariância ($p = 700$) é aparente imediatamente ao observarmos o posto, o condicionamento e se a matriz estimada é positiva definida. Para o estimador empírico, a matriz $\hat{\boldsymbol{S}}$ apresenta posto igual a 22. O condicionamento é infinito, visto a matriz ser singular, além da estimativa não ser positiva definida, pois apenas 22 autovalores são positivos. Para o estimador por encolhimento, os resultados apresentam melhor comportamento e características muito mais desejáveis, visto que o posto é igual 700, o condicionamento é igual a aproximadamente 423.000, e a matriz obtida é positiva definida, com todos autovalores positivos.

Com os mesmos dados foram estimadas as matrizes de correlação empírica e por encolhimento. Para ambas as matrizes os sinais das correlações são exatamente iguais, uma garantia de que o estimador por encolhimento não causa alteração na direção da correlação, visto que o procedimento é apenas de encolher os elementos em direção à matriz identidade. Em geral, as estimativas de correlação empíricas apresentam uma faixa dinâmica maior (em geral são sobre-estimadas) que as obtidas por encolhimento, devido à maior variância do estimador empírico. Esse comportamento pode ser observado na Fig. 9. Vê-se na figura que, as correlações estimadas por encolhimento, naturalmente se situam no intervalo $[-0.4, +0.4]$, enquanto uma grande parcela das correlações para este conjunto de genes no estimador empírico apresenta valor absoluto maior que $\rho = 0.5$. É possível comparar estes histogramas com os histogramas de correlação globais na base de dados TBDB (TB Database, 2015b). Para o gene apresentado na Fig. 9, o

[1] um instante amostral contendo apenas duas repetições.

histograma global (determinado com um conjunto de $n = 1260$ amostras[2]) obtido da base TBDB está mostrado na Fig. 10 e apresenta correlações essencialmente na mesma faixa das correlações obtidas com o estimador por encolhimento. Nota-se aqui também a vantagem da utilização do estimador por encolhimento, o qual reduz a tendência à super-estimação das correlações que ocorre no estimador empírico. Exemplos das correlações obtidas para outros genes do conjunto estão apresentados nas últimas três figuras do capítulo, junto com os histogramas de correlações globais correspondentes obtidos da base TBDB. Em todas as instâncias observa-se uma melhor aproximação das correlações com o uso do estimador por encolhimento.

2.4 CONCLUSÃO

Neste capítulo foi abordado o tema da estimação da matrizes de covariância em problemas da classe *p-grande, n-pequeno*. Um exemplo típico dessa classe é a análise de dados biológicos de transcriptômica, ou expressão genética. Medidas de expressão genética de todo o conjunto de genes de um organismo, i.e. medidas em escala genômica, podem ser obtidas com relativa facilidade com compreensivos experimentos de micromatrizes de DNA. Essa técnica experimental permite obter de uma só vez a medida de expressão dos milhares de genes que compõem o genoma de uma célula. No caso do *Mycobacterium tuberculosis*, o número de genes identificados ultrapassa 4000. Tais experimentos, embora atualmente simples de realizar, são caros, sendo o custo o principal fator limitante para produzir várias repetições de um experimento. Assim, a alta dimensionalidade dos dados associada ao baixo número de observações torna todos os problemas de análise transcripcional, em geral, problemas do tipo *p-grande, n-pequeno*, com todas as dificuldades e desafios estatísticos associados. Um desses principais desafios é a estimação da matriz de covariância dos dados.

Em muitos dos trabalhos publicados na área de bioinformática (SCHÄFER; STRIMMER, 2005), o estimador frequentemente utilizado é o estimador empírico não-polarizado da Eq. (2.2). Mostramos que esse estimador é altamente ineficiente e mesmo indesejável em situações de poucas amostras de alta dimensionalidade. Os problemas, nesse caso, são a sua alta variância, alta magnitude de erro quadrático médio (i.e. norma-F da diferença), mau condicionamento, apresentar autovalores negativos quando estimados em precisão finita (i.e. não produzindo

[2]Dados das publicações (BOSHOFF et al., 2004) e (VOSKUIL et al., 2004)

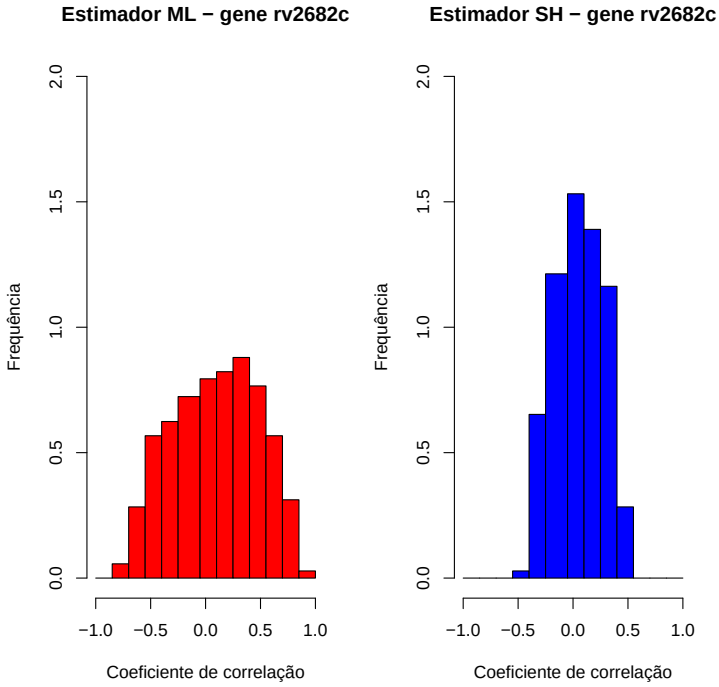

Figura 9 – Histograma das correlações do gene Rv2682c com todos os outros genes do conjunto. Em vermelho os resultados obtidos com o estimador empírico. Em azul as correlações estimadas por encolhimento. Para esta simulação $p = 700$ e $n = 23$. Dados de exposição do Mtb à mefloquina. Além da matriz estimada por encolhimento corrigir diversos problemas do estimador empírico como: autovalores todos positivos, bom condicionamento e posto completo, para genes individuais, o estimador $\boldsymbol{S}^*$ também corrige a tendência do estimador empírico a super-estimar os valores das correlações. *Fonte: Elaborada pelo autor.*

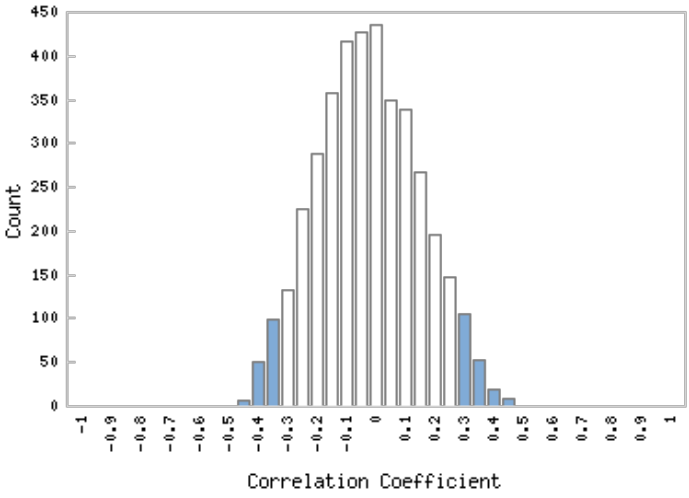

Figura 10 – Histograma das correlações globais (inclui todos os genes do Mtb). *Fonte: TBDB (TB Database, 2015b).* O link específico para esta imagem é o da página das correlações em `http://tuberculosis.bu.edu/tbdb_sysbio/CC/Rv2682c.html`.

uma estimativa positiva definida) e posto incompleto, resultando em matrizes singulares e portanto não invertíveis.

Foi apresentado o estimador por encolhimento (*shrinkage estimator*), o qual corrige todos os problemas do estimador empírico identificados, sem incorrer em um significativo aumento na complexidade computacional. Esse pequeno aumento na complexidade computacional é decorrente do fato que esse estimador pode ser determinado analiticamente, uma grande vantagem em relação a outros estimadores assintóticos como *bootstrap* e MCMC.

Em (SCHÄFER; STRIMMER, 2005), os autores chamam a atenção para a baixa popularidade desse estimador em trabalhos da área, visto as diversas vantagens que apresenta em relação ao estimador empírico. O estimador por encolhimento apresenta menor erro quadrático médio, menor variância, além de produzir estimativas com bom condicionamento e posto completo mesmo para número de observações 10 vezes menor que a quantidade de amostras.

Finalizamos o capítulo com a estimação das matrizes de covariância a partir de dados reais, avaliando os histogramas de correlação e correlação parciais dos parâmetros estimados. Desse estudo conclui-se que:

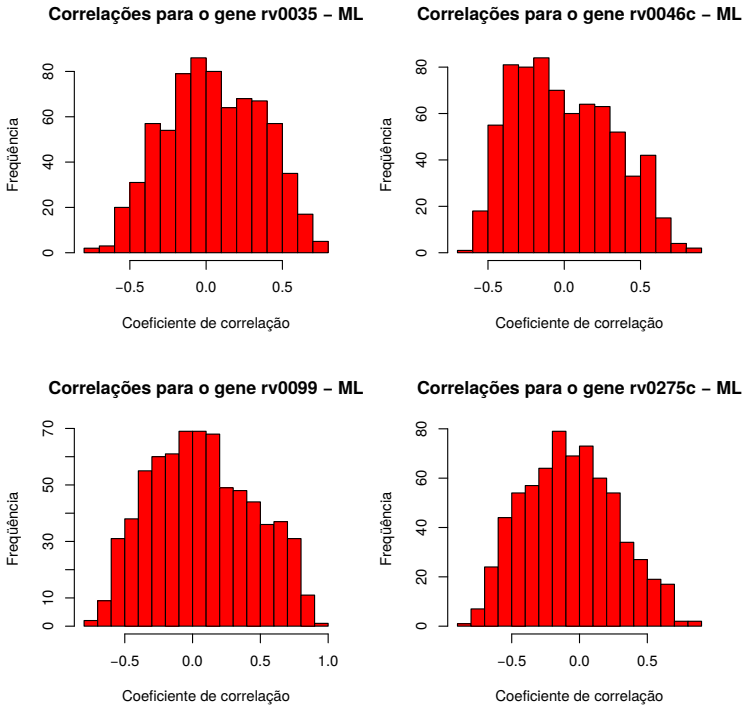

Figura 11 – Histogramas das correlações de cada gene indicado com todos os outros genes do conjunto. Os resultados apresentados são os obtidos com o estimador empírico. *Fonte: Elaborada pelo autor.*

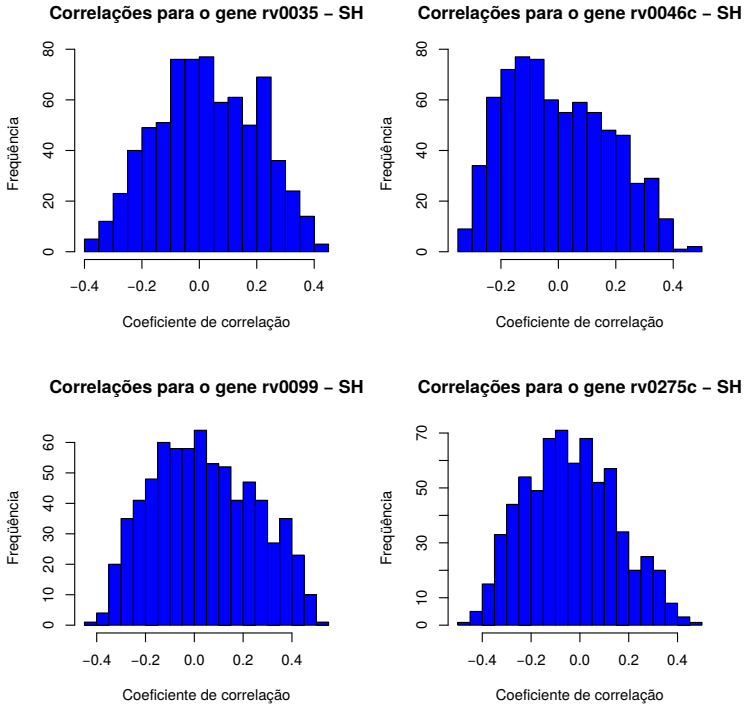

Figura 12 – Histogramas das correlações de cada gene indicado com todos os outros genes do conjunto.Os resultados apresentados são os obtidos com o estimador por encolhimento. *Fonte: Elaborada pelo autor.*

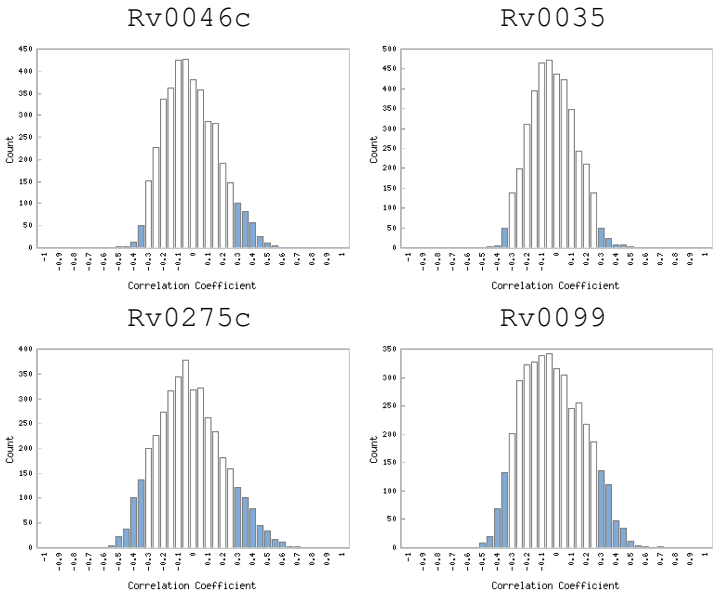

Figura 13 – Histogramas das correlações globais para os genes *Rv0046c*, *Rv0035*, *Rv0275c* e *Rv0099*. *Fonte: TBDB (TB Database, 2015b).* Os links específicos para as imagens encontram-se em `http://tuberculosis.bu.edu/tbdb_sysbio/CC/XXXXXX.html`, substituindo XXXXXX pelo nome do gene desejado.

- o estimador empírico não-polarizado, apesar de sua popularidade, não deve ser utilizado em problemas *p-grande,n-pequeno*, ou seja, quando $n \ll p$;
- o estimador por encolhimento é um estimador bastante eficiente e portanto um bom substituto em problemas $n \ll p$;
- o estimador por encolhimento apresenta características bastante atraentes, como garantia de uma estimativa positiva definida e bom condicionamento;
- em termos médios, o estimador por encolhimento apresenta menor erro quadrático médio e menor variância que o estimador empírico.

3 FLUXOS METABÓLICOS

Para que seja possível estudar a fisiologia e metabolismo de um organismo, ou mesmo para estimar um modelo de predição para o metabolismo, é necessário que tenhamos dados de *fluxos metabólicos*. Medições de fluxo metabólico, todavia, são difíceis de realizar (ZUPKE; SINSKEY; STEPHANOPOULOS, 1995), e no estado atual da tecnologia (WITTMAN, 2007) nem todos os fluxos que compõem o fluxoma podem ser medidos. Para contornar tal dificuldade, utiliza-se uma técnica já bem aceita e estabelecida, a qual permite a obtenção de dados de fluxo metabólico a partir da matriz estequiométrica das reações metabólicas de um organismo em um procedimento de otimização. Essa técnica, chamada de análise de balanço de fluxos ou FBA (*flux balance analysis*), é um método especialmente popular para estimação do metabolismo de organismos procarióticos sob diferentes condições ambientais, como hipoxia ou escassez de nutrientes. O uso do FBA, visto este ser um procedimento de otimização, requer a especificação de uma função objetivo que represente o objetivo metabólico da célula para cada condição ambiental. Várias funções objetivo já foram propostas (SCHUETZ; KUEPFER; SAUER, 2007) e podem ser utilizadas como suposto objetivo metabólico. Algumas funções objetivo resultam em problemas de otimização linear, ao passo que outras resultam na formulação de um problema de otimização quadrática.

A técnica de FBA é construída em cima da informação fornecida por uma rede metabólica *in silico* (como, por exemplo, a apresentada em (BESTE et al., 2007)) e produz como saída um vetor de fluxos metabólicos $\boldsymbol{v}$ para as reações da rede. Esse vetor maximiza a função objetivo definida pelo investigador.

A determinação da função objetivo é um tema complexo e ainda sem solução, porém é fato aceito que a eficácia e precisão das predições dos fluxos é dependente da escolha da função objetivo, que por sua vez pode ser dependente da condição ambiental. Uma função objetivo normalmente utilizada para estudo de um organismo em homeostase é a *equação de síntese de biomassa*, a qual representa a geração de biomassa pela célula, ou em outras palavras, o crescimento celular. Essa expressão é definida como uma reação metabólica de conveniência (que não existe realmente na célula) que leva em conta as proporções de precursores metabólicos e macromoléculas necessárias para o crescimento celular de um determinado tipo de organismo. A função de biomassa é útil em uma grande variedade de problemas e diversas inferências bio-

logicamente validadas foram obtidas com esse objetivo (VARMA; PALSSON, 1994).

Entretanto, essa função objetivo não é necessariamente a melhor escolha em casos nos quais não podemos assumir que o objetivo metabólico da célula seja crescimento. No caso de bactérias como o Mtb, um dos objetivos do estudo dos fluxos metabólicos seria colocar a célula em uma situação de estresse fisiológico, expondo-a a um composto bactericida e observar o comportamento metabólico de sobrevivência da célula. Se estudamos a célula com FBA a partir da maximização de biomassa, estamos assumindo um objetivo celular incorreto, visto que o foco de todas as capacidades metabólicas do organismo está na tentativa de sobrevivência. Em situações extremas o sistema de regulação da célula provavelmente desliga todos circuitos metabólicos que não são essenciais à sua sobrevivência, concentrando sua atividade em mecanismos de detoxificação e recuperação de áreas celulares danificadas pelo composto bactericida, como no caso por exemplo de drogas que interrompem a síntese da parede celular.

Nesta seção o nosso objetivo é apresentar uma nova maneira de definir a função objetivo para FBA em casos quando a maximização de biomassa não pode ser assumida. Assim, apresentaremos a seguir os conceitos e informações básicas para entendimento da proposta. Explicaremos sobre a determinação e construção da rede metabólica *in silico*, os procedimentos experimentais de proteômica e análise de balanço de fluxos. Esses três temas serão tratados de forma breve e mais detalhes podem ser encontrados nas diversas referências indicadas na bibliografia.

3.1 CONSTRUÇÃO DA REDE METABÓLICA

Para uma melhor compreensão do método FBA e consequentemente da proposta deste capítulo, é importante esclarecer os conceitos de reação metabólica, metabólito, estequiometria e fluxo. Diversas referências podem ser consultadas para explicações mais detalhadas sobre estes e outros conceitos relacionados. No caso de microorganismos procariotas como o *Mycobacterium tuberculosis*, ver por exemplo (MADIGAN et al., 2006), e para organismos multicelulares (ALBERTS et al., 2002).

Uma reação bioquímica metabólica é um processo que transforma um conjunto de substratos (entradas da reação) em um outro conjunto de produtos (saídas da reação), como no exemplo abaixo de

descarboxilação (eliminação de CO_2) do ácido glutâmico:

$$C_5H_9NO_4 \xrightarrow{R_1} C_4H_9NO_2 + CO_2$$

Tanto os substratos quanto os produtos são chamados *metabólitos*. Os metabólitos são todas as moléculas e compostos orgânicos presentes no ambiente celular que são necessários para o funcionamento do organismo, tais como moléculas de água, hidrogênio, oxigênio, lipídeos, adenosina tri-fosfato (ATP), glicose, entre outros.

Em reações metabólicas, a quantidade de cada substrato/produto é dada em moles, e na formula da reação o número de moles é indicado como um *coeficiente estequiométrico*, anteposto à fórmula estrutural do metabólito. No exemplo a seguir, temos a reação metabólica conveniência de produção biomassa para o Mtb *in vitro*:

$$\begin{aligned}
&0.214\ \text{PROTEÍNA} \\
+\ &0.036\ \text{RNA} \\
+\ &0.022\ \text{DNA} \\
+\ &0.050\ \text{MICROMOLECULAS} \\
+\ &0.006\ \text{PE} \\
+\ &0.016\ \text{TAGbio} \\
+\ &0.040\ \text{PIMs} \\
+\ &0.186\ \text{LAM} \\
+\ &0.208\ \text{MAPC} \\
+\ &0.035\ \text{P-L-GLX} \\
+\ &0.007\ \text{CL} \\
+\ &0.054\ \text{LM} \\
+\ &47\ \text{ATP} \\
=\ &1.0\ \text{BIOMASSA} + 47\ \text{ADP} + 47\ \text{PI}
\end{aligned} \tag{3.1}$$

Essa reação[1] foi compilada em (BESTE et al., 2007) e é frequentemente utilizada em FBA como objetivo de maximização. Os experimentos

[1] Os nomes dos metabólitos presentes nessa equação estequiométrica seguem a nomenclatura abreviada usada na construção do modelo metabólico GSMN-TB em (BESTE et al., 2007) e um glossário pode ser encontrado no material suplementar 6 do respectivo artigo. Aqui, **PE** (L-1-phosphatidyl-ethanolamine), **TAGbio** (triacylglycerol), **PIMs** (phosphatidylinositol mannosides), **LAM** (lipoarabinomannan),

e cálculos usados na determinação dessa expressão são complexos e não são o foco deste trabalho, porém estão disponíveis e podem ser encontrados em detalhes no material suplementar de (BESTE et al., 2007). O importante para a nossa exposição é que os coeficientes estequiométricos indicam uma estimativa da quantidade de moles de cada composto e macromolécula necessários para a produção de um mole de biomassa para o Mtb. Nota-se que, além de biomassa, são produzidos nessa reação 47 moles de ortofosfato e adenosina-difosfato (ADP) resultantes da hidrólise das respectivas moléculas de ATP usadas como fonte energética.

Na proposta deste capítulo é nosso intuito mostrar que em situações de exposição a compostos bactericidas, i.e. em situações de sobrevivência, essa função objetivo não deveria ser utilizada, mas uma outra que melhor represente o objetivo metabólico da célula.

Na reação anterior de descarboxilação do ácido glutâmico, os coeficientes não são colocados pois todos são iguais à unidade, i.e. exatamente 1 mole de $C_5H_9NO_4$ produz 1 mole de $C_4H_9NO_2$ e 1 mole de dióxido de carbono.

A *estequiometria* das reações garante que todos os substratos sejam completamente transformados em produtos, sem que nenhuma molécula se perca ou seja sintetizada, i.e. todas as reações são *balanceadas* e, portanto, é válida a Lei da Conservação de Massa.

A taxa (velocidade) de conversão de moléculas de substrato em moléculas de produtos é dada pela medida de *fluxo metabólico*. O fluxo de reações bioquímicas que ocorrem de forma espontânea no ambiente celular são, em geral, lentas. Todavia, as demandas metabólicas dos organismos são grandes e rápidas, e exigem um processamento com fluxo maior.

Para entender como a célula soluciona o problema e aumenta o fluxo metabólico das reações, é importante compreender conceitos básicos sobre o papel da energia nas reações.

De forma simplificada, todas as reações químicas estão associadas com *mudanças de energia*. Algumas reações requerem energia, enquanto outras produzem energia. Nas reações metabólicas não temos interesse na energia despendida na forma de calor, mas sim na energia livre (G), que é a energia disponível para realizar trabalho dentro da célula. A variação em energia livre é indicada pelo símbolo ΔG°, quando assume-se que essa variação ocorre em condições padrão de temperatura (25°C), pressão (1 atm), pH (7.0) e concentração (1 mole). Se

---

MAPC (mycolic-acid-arabinogalactan-peptidoglycan complex), **P-L-GLX** (poly-L-glutamate-glutamine), **CL** (cardiolipin), **LM** (lipomannan), **PI** (phosphate)

uma reação produz energia o seu ΔG° será negativo e a reação é designada *exergônica* e a energia livre resultante será armazenada dentro da célula na forma de ATP. Se $\Delta G^\circ > 0$, então a reação necessita de energia livre para se concretizar, e tal reação é dita *endergônica*.

Dentro da célula não basta que duas moléculas estejam próximas para que reajam, é necessário que a reação entre elas seja exergônica, liberando energia, pois de outra forma a reação não ocorre espontaneamente. Além da variação de energia, é necessário que as moléculas ultrapassem um limite energético mínimo conhecido como *energia de ativação*, sem o que a reação não ocorre. Percebe-se que para que as reações ocorram mais facilmente e com um fluxo mais alto, é necessário que a célula auxilie o processo, e isso é feito diminuindo a energia necessária para que uma reação ocorra. As enzimas são proteínas que *diminuem* a quantidade de energia de ativação necessária para uma reação, e esse processo é denominado *catálise*. A presença da enzima altera apenas a taxa de processamento (i.e. fluxo) da reação. Ela não pode alterar o delta de energia de uma reação tornando, por exemplo, uma reação endergônica em uma reação exergônica. Em geral, as enzimas atuam sobre as reações exergônicas (que não necessitam de energia externa), e o ganho em fluxo para certas reações pode ficar na faixa de 10^4 vezes.

A cada reação metabólica corresponde uma enzima, pois essas são altamente específicas, e raramente catalisam mais de uma reação diferente. Todas as enzimas são reversíveis e podem catalisar tanto a reação direta quanto a reação reversa. Porém, no caso de reações fortemente exergônicas ou fortemente endergônicas, o trabalho da enzima em geral é unidirecional.

Portanto, sempre que a célula necessite de uma determinada reação metabólica com maior fluxo, a enzima correspondente deve estar presente em quantidade suficiente no citoplasma, ou então deve ser sintetizada *de novo* a partir do gene correspondente no DNA. É importante notar que existe uma relação muito próxima entre a quantidade de enzima presente no ambiente celular e o fluxo da reação correspondente. Essa é a suposição principal subjacente à ideia apresentada neste capítulo.

Como vimos, para o método FBA é necessário construir a *rede metabólica* do organismo cujo metabolismo está sendo estudado. As reações bioquímicas metabólicas, visto que atuam sobre o mesmo conjunto de metabólitos, podem ser agrupadas em sequência ou de forma paralela, formando uma complexa rede de processamento de metabólitos (NEWMAN, 2010). A construção da rede metabólica é determinada

por compilação das reações individuais que compõem processos metabólicos conhecidos do organismo ou por comparação ortológica. Uma referência com protocolos específicos e diretrizes para compilação de redes metabólicas está disponível em (THIELE; PALSSON, 2010). Para as simulações deste capítulo, que foram realizadas para o organismo Mtb, a rede metabólica utilizada é uma versão corrigida e atualizada da rede GSMN-TB (*genome-scale metabolic network - TB*) apresentada em (BESTE et al., 2007) (Para maiores detalhes, ver (MONTEZANO et al., 2015)).

Redes metabólicas já foram brevemente mencionadas na introdução do trabalho, e estão disponíveis para um número cada vez maior de organismos e com acesso público através de bases de dados como KEGG (Kanehisa Labs, 2015) (para uma gama de diversos organismos) ou o TBDB (TB Database, 2015b), no caso de uma base específica da bactéria Mtb.

3.2 PROTEÔMICA

Como será apresentado em breve, a proposta deste capítulo é definir a função objetivo para FBA a partir de dados experimentais. A escolha de utilizar dados de expressão proteica será discutida a seguir.

Quando falamos sobre a construção da rede metabólica, vimos que o fluxo metabólico das reações é dependente da presença de uma proteína catalisadora, chamada enzima. Na definição da função objetivo estamos assumindo implicitamente que a quantidade expressa de uma determinada enzima na célula é proporcional ao fluxo metabólico através da reação catalisada por tal enzima. Isto é, assume-se que a concentração enzimática é um bom indicativo da velocidade da reação catalisada correspondente. É certo que essa suposição serve para muitos casos, porém vale citar que ela não compreende as seguintes situações:

- a enzima está expressa na célula, porém não está atuante devido a um determinado metabólito não estar disponível no citoplasma;
- a enzima está expressa na célula, porém não está na sua conformação tridimensional ativa, por exemplo, quando ainda não estiver fosforilada;
- a enzima está disponível na célula, porém uma determinada coenzima necessária à reação não está presente no citoplasma;
- o nível basal da proteína é alto, independentemente do seu estado

de ativação;

- existem enzimas que não são necessárias à célula e estão em processo de degradação lento quando as medições são feitas.

Todavia, ainda que essa suposição não leve em conta os casos acima, os quais não podem ser resolvidos apenas observando os dados de proteômica, a suposição de equivalência entre expressão proteica e fluxo metabólico é útil, e é nosso intuito mostrar que predições feitas com FBA utilizando essa suposição para definir um novo tipo de função objetivo em situações de estresse são mais corretas do que as predições obtidas com a função objetivo de biomassa.

Portanto, assumindo que os dados de proteômica sejam a escolha certa, é necessário obter tais medidas de expressão de uma amostra celular. Um experimento de proteômica mede o *proteoma* celular, i.e. o conteúdo proteico presente no citoplasma celular em um determinado instante. Essas medidas também são chamadas de *medidas de expressão genética em nível proteico*, um conceito que deixa mais claro o fato que essa é uma medida de atividade pós-traducional, i.e. ela mede o quanto de mRNA (transcriptoma) realmente se transforma em proteínas. As medições do proteoma são muito úteis, pois podem indicar (assim como o transcriptoma, porém de forma mais precisa) alterações na composição proteica causada por doenças ou pela ação de drogas. Por sua vez, associando essas alterações nas proteínas com alterações nos respectivos fluxos metabólicos, acreditamos ser possível melhorar as predições sobre o metabolismo obtidas com FBA.

Os experimentos de proteômica apresentam alta relação sinal-ruído, diferentemente dos experimentos com micromatrizes de DNA, os quais sofrem maior degradação devido ao ruído intrínseco ao processo de medição. A vantagem do experimento de micromatriz de DNA é ser mais abrangente que um experimento similar de proteômica, como explicaremos a seguir.

A medição do proteoma é realizada em duas partes:

- separação das proteínas presentes em uma amostra por eletroforese (TÖZEREN; BYERS, 2004);
- contagem e identificação das proteínas por espectrometria de massa (PAVIA et al., 2010).

A separação das proteínas é feita com base no seu tamanho e na sua carga. A separação é feita sobre uma placa com gel que limita a velocidade de movimentação das proteínas, criando um padrão bidimensional que pode ser usado para identificar visualmente alterações

no proteoma (LESK, 2007). Caso seja necessário identificar proteínas individuais, os grupos sobre o gel podem ser extraídos, digeridos e identificados por espectrometria de massa, e por comparação com uma base de dados de proteínas. Embora a exatidão do método seja alta para todas as proteínas identificadas, há limitações no processo, pois proteínas que não possam ser digeridas, sejam muito pequenas ou não estejam presentes na base de dados, não serão encontradas.

Para as simulações que serão apresentadas, os dados de proteômica foram obtidos de experimentos com o *Mycobacterium tuberculosis* exposto ao composto bactericida mefloquina. Maiores informações e detalhes sobre os procedimentos experimentais do conjunto de dados utilizado nas simulações podem ser obtidos em (MONTEZANO et al., 2015).

3.3 ANÁLISE DE BALANÇO DE FLUXOS

O procedimento de análise de balanço de fluxos (*flux balance analysis* - FBA) é um método popular para estudo do metabolismo. Material introdutório sobre a técnica de programação linear subjacente ao FBA, ou explicações sobre sua aplicação ao metabolismo podem ser obtidas em diversas publicações disponíveis (PAPADIMITRIOU; STEIGLITZ, 1982; ORTH; THIELE; PALSSON, 2010; RAMAN; CHANDRA, 2009; VARMA; BOESCH; PALSSON, 1993; VARMA; PALSSON, 1994; SCHUETZ; KUEPFER; SAUER, 2007; LAKSHMANAN et al., 2012; COLIJN et al., 2009). Apresentamos a seguir alguns conceitos introdutórios relevantes para uma melhor compreensão das implicações da nossa proposta de uma nova função objetivo.

FBA é um método de otimização linear aplicado à análise do metabolismo, em que unicamente utiliza-se o conhecimento estequiométrico da rede de reações metabólicas do organismo. A aplicação do método requer algumas informações prévias sobre o problema:

- conhecimento da rede metabólica do organismo;
- definição das restrições de fluxo para cada reação da rede (determinando *capacidade* e *reversibilidade*);
- definição de uma função objetivo biologicamente adequada ao problema, a qual será utilizada como alvo do processo de otimização.

A formulação do FBA é dada pelo seguinte problema de pro-

gramação linear (PAPADIMITRIOU; STEIGLITZ, 1982):

$$\begin{array}{lll} \min\limits_{\boldsymbol{v}} & f(\boldsymbol{v}) \triangleq \boldsymbol{c}^T\boldsymbol{v} & \boldsymbol{v} \in \Re^n \\ \text{s. a} & \boldsymbol{S}\boldsymbol{v} = 0, & \boldsymbol{v}_{min} \leq \boldsymbol{v} \leq \boldsymbol{v}_{max} \end{array} \tag{3.2}$$

em que $\boldsymbol{S}$ é a matriz estequiométrica representativa da rede metabólica, $\boldsymbol{v} = [v_1, v_2, \ldots, v_q]^T$ é o vetor de fluxos com dimensão q igual ao número de reações metabólicas que compõem a rede *in silico*, e $\boldsymbol{c}$ é o vetor de custos, uma terminologia da programação linear (PAPADIMITRIOU; STEIGLITZ, 1982) empregada em FBA para indicar os coeficientes da combinação linear de fluxos que define a função objetivo do problema. Embora o problema definido na Eq. (3.2) indique uma minimização, transformá-lo em um problema de maximização é trivial, bastando inverter o sinal da função objetivo.

O problema de otimização de FBA requer a especificação de restrições lineares de igualdade e restrições lineares de desigualdade que limitam os valores mínimo e máximo para cada fluxo. Enquanto as primeiras são restrições *estequiométricas*, que garantem uma distribuição de fluxos balanceada (i.e. em equilíbrio), as últimas são restrições de capacidade ou termodinâmicas, que definem a reversibilidade ou irreversibilidade de cada fluxo.

Uma vez construído o modelo metabólico do organismo representado matematicamente pela matriz estequiométrica $\boldsymbol{S}$, e limitado o espaço de otimização pelas restrições de capacidade e reversibilidade, pode-se calcular uma distribuição ótima de fluxos metabólicos $\boldsymbol{v}^*$. Esse vetor ótimo de fluxos deve ser representativo do comportamento metabólico que é assumido pelo organismo para a função objetivo $f(\boldsymbol{v})$ selecionada.

Na Fig. 14 apresentamos um diagrama esquemático do procedimento de otimização FBA descrito acima. Todas as restrições do método FBA são lineares, de igualdade ou desigualdade. A função objetivo é uma combinação linear dos fluxos metabólicos do vetor $\boldsymbol{v}$. O objetivo deste capítulo é apresentar um novo método para definir a combinação linear de fluxos que será usada como função objetivo do processo de otimização em situações específicas de sobrevivência (bloco indicado em vermelho na Figura 14).

Para uma bactéria como o Mtb, definindo uma função objetivo de biomassa, o resultado de maximizar esse objetivo será um vetor de fluxos $\boldsymbol{v}^*$ que suporta o crescimento bacterial como objetivo final da célula.

Entretanto, é fato que crescimento e reprodução deixam de ser

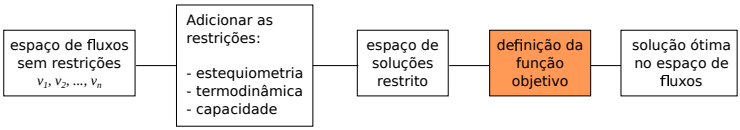

Figura 14 – Diagrama em blocos esquemático do procedimento FBA. FBA é um processo de otimização que busca uma configuração metabólica ótima no espaço de fluxos metabólicos. Esse espaço, que inicialmente é irrestrito, é limitado pelas informações da matriz estequiométrica $\boldsymbol{S}$ e restrições de reversibilidade e capacidade. O algoritmo de otimização busca nesse espaço com restrições uma configuração de fluxos metabólicos que maximize uma determinada função objetivo. Em FBA, a função objetivo é sempre uma combinação linear de fluxos metabólicos (variáveis do espaço). A configuração metabólica ótima é o vetor de fluxos metabólicos ótimo $\boldsymbol{v}^*$, o qual satisfaz todas as restrições impostas e maximiza o valor da função objetivo $f(\boldsymbol{v})$. *Fonte: Elaborada pelo autor.*

os objetivos primeiros de bactérias expostas a compostos bactericidas, dando lugar a um objetivo mais urgente: sobreviver. Como mostra a Figura 15, quando as bactérias não são sujeitas a nenhum tipo de estresse biológico, o funcionamento metabólico desenrola-se normalmente (situação A) utilizando os nutrientes disponíveis na produção de energia, sendo que as colônias de bactérias perseguem dois principais objetivos: crescimento (aumento de biomassa) e reprodução. É fácil perceber que, nesse caso, uma análise com FBA pode ser realizada utilizando como função objetivo metabólico a produção ótima de biomassa, visto que esse é realmente o objetivo buscado pelo organismo.

Todavia, quando o organismo é exposto a um composto antibiótico (situação B), não podemos aceitar que o foco do metabolismo continue a ser a produção de biomassa e a reprodução, pois o organismo necessita lidar com uma situação mais urgente. Nesse ponto (situação C), as diversas funções metabólicas necessárias na produção de biomassa são desativadas ou relegadas a um plano secundário enquanto a célula passa por um processo de reprogramação celular que a permita lidar com a nova situação. Ainda é possível estudar essa situação metabólica alternativa com FBA, porém não é mais possível assumir que o objetivo metabólico da célula ainda seja a maximização de biomassa. Um nova proposta de função objetivo é necessária.

Dessa forma, a pergunta deste capítulo é: Que função objetivo deve ser usada se queremos estudar o metabolismo da bactéria exposta a agentes bactericidas? No restante do capítulo propomos uma resposta possível para essa pergunta, a qual é de fácil implementação e

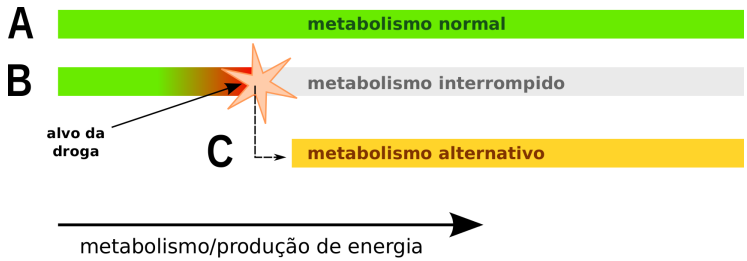

Figura 15 – Em FBA, é comum o uso de uma função objetivo que maximize a produção de biomassa do organismo, indicando que em condições normais de desenvolvimento (situação A) o objetivo do organismo é crescimento e reprodução. Entretanto, quando o organismo é exposto a um composto antibiótico e sua sobrevivência está ameaçada (situação B), a célula será reprogramada de forma a apresentar um metabolismo alternativo que maximize as suas chances de sobrevivência. Nesse momento (situação C), não é possível estudar o metabolismo do organismo com FBA ainda assumindo que o seu objetivo é produção ótima de biomassa e uma função objetivo alternativa é necessária. *Fonte: Elaborada pelo autor.*

leva a resultados promissores quando comparada a outras possibilidades existentes.

Muitos estudos têm tentado de diversas maneiras introduzir dados experimentais na técnica original de forma a 'guiar' melhor as predições de FBA (COVERT; SCHILLING; PALSSON, 2001; SHLOMI et al., 2008; COLIJN et al., 2009; YIZHAK et al., 2010; COVERT et al., 2004; CHANDRASEKARAN; PRICE, 2010). Os dados são geralmente incorporados na forma de restrições, e quando ambos transcriptômica e proteômica são utilizados, o objetivo é resolver inconsistências entre as medidas de atividade transcripcional e traducional. A proposta deste capítulo é apresentar uma maneira simples de incluir dados experimentais em FBA via definição da função objetivo.

Um recente estudo (MACHADO; HERRGARD, 2014) comparou diversos desses métodos com resultados pouco animadores, visto que em muitos casos os resultados não são melhores, podendo ser mesmo piores e mais distantes das medidas experimentais do que os obtidos com a técnica FBA sem o uso de dados experimentais. Um dos métodos que apresentou o melhor desempenho nas várias simulações realizadas foi o método E-flux, descrito em (COLIJN et al., 2009). Assim, selecionou-se esse método para realizar comparações com os resultados obtidos com a nossa proposta, a ser apresentada na próxima seção.

3.4 PROTEÍNAS DE SOBREVIVÊNCIA

Nesta seção descrevemos essa nova proposta de função objetivo, mostrando como dados de proteômica podem ser usados para determinar um objetivo para FBA em situações em que o organismo tenha sido exposto a um antibiótico e esteja sob estresse metabólico, de forma que não seja adequado assumir crescimento ótimo como objetivo metabólico da célula. A forma da função objetivo é a de uma combinação linear de fluxos, como apresentado em diversos trabalhos como (COLIJN et al., 2009; RAMAN; RAJAGOPALAN; CHANDRA, 2005). Entretanto, os coeficientes da combinação linear são determinados de maneira diversa. A forma da função objetivo é dada por:

$$f(\boldsymbol{v}) = \sum_{i=1}^{q} c_i v_i = \boldsymbol{c}^T \boldsymbol{v} \tag{3.3}$$

em que $\boldsymbol{v} = [v_1, v_2, \ldots, v_q]^T$ é o vetor de fluxos e $\boldsymbol{c}$ é o vetor de coeficientes da combinação linear que será maximizada. A dimensão q de ambos os vetores é igual ao número total de reações (i.e. fluxos metabólicos) no modelo *in silico* utilizado para a rede metabólica. No caso de maximização de biomassa, o vetor $\boldsymbol{c}$ é um vetor com zero em todas as posições, exceto pelo fluxo correspondente à produção de biomassa:

$$f_{bio}(\boldsymbol{v}) = \boldsymbol{c}^T \boldsymbol{v} = [1, 0, 0, \ldots, 0] \begin{bmatrix} v_{bio} \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} \tag{3.4}$$

em que v_{bio} é o fluxo de produção de biomassa definido no modelo da rede metabólica, e que representa a expressão definida na Eq. (3.1). No método apresentado, ainda será usada a mesma combinação linear de fluxos, contudo ao invés de maximização de biomassa, os coeficientes c_i são calculados a partir de dados de proteômica. A Figura 16 apresenta uma visão esquemática da proposta. A matriz estequiométrica é obtida a partir das informações da rede metabólica, enquanto as restrições individuais para cada fluxo são determinadas a partir de considerações termodinâmicas e de capacidade para cada reação. A função objetivo é determinada a partir de dados experimentais como apresentado a seguir.

Antes de explicarmos o método proposto, descrevemos a seguir

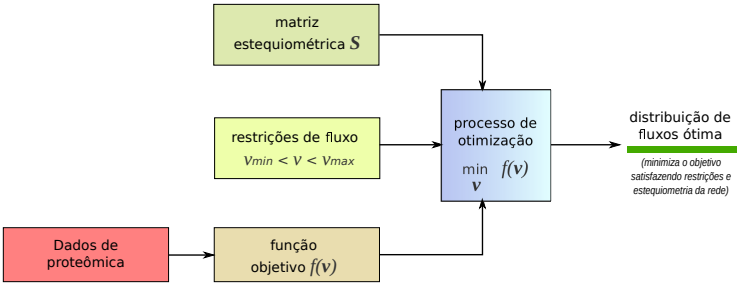

Figura 16 – FBA é um método de otimização linear com restrições. A matriz estequiométrica é obtida a partir de dados biológicos disponíveis em bases públicas sobre o metabolismo, contendo a relação entre metabólitos e reações metabólicas. As restrições para os fluxos são definidas a partir de conhecimento biológico sobre o comportamento termodinâmico da reação correspondente e sua capacidade. No método proposto, a função objetivo é definida como uma combinação linear de fluxos catalisados por enzimas que estão presentes no conjunto de dados experimentais. Assume-se que a maximização dos fluxos dessas proteínas é mais próximo do real objetivo metabólico da célula do que a maximização de biomassa. *Fonte: Elaborada pelo autor.*

os experimentos realizados para a obtenção dos dados de proteômica. Todos os experimentos para obtenção de dados de transcriptômica e proteômica utilizados neste trabalho foram projetados e realizados pelo Departamento de Microbiologia da Universidade do Estado do Oregon, sob a coordenação e supervisão do Prof. Luiz E. Bermudez. Para o estudo deste capítulo, foram utilizados dados experimentais de proteômica do Mtb em um tratamento com mefloquina e uma condição de controle em três instantes amostrais. As culturas de bactérias foram obtidas essencialmente como descrito em (MCNAMARA et al., 2012, 2013). A cepa de bactérias utilizada nos experimentos foi o *Mycobacterium tuberculosis* H37Rv obtida de ATCC[2]. As bactérias foram cultivadas sobre meio ágar Middlebrook 7H-10 enriquecido com ADC[3]. Para análise de proteômica as colônias de *Mycobacterium tuberculosis* foram cultivadas em um volume de 100 ml em frascos de 500 ml por 7 dias com uma densidade de células de 1×10^8 cél./ml. Na condição de controle foi utilizado o solvente DMSO e para as condições

[2] American Type Culture Collection, Manassas, Virginia, 20110, USA.
[3] *ADC*: albumina, dextrose e catalase.

de tratamento[4] [5] foram utilizados 8μg/ml de mefloquina dissolvidos em DMSO. Culturas foram subsequentemente separadas por centrifugação a 3000 r.p.m. por 10 minutos em intervalos de 6 horas, 2 e 4 dias. As células foram lavadas 3 vezes com solução salina e Tween (0.8% m/v NaCl/ 0.05 v/v Tween 80)[6] [7]. Os aglomerados de células foram colocados novamente em suspensão em uma mistura para lise celular das bactérias, contendo 80 μl de inibidor de protease. As bactérias foram dissolvidas por agitação (mini BeadBeater com velocidade 8 por 2 minutos) e mantidas no gelo por um número adicional de minutos entre cada sessão.

As amostras foram centrifugadas a 15000 r.p.m. a uma temperatura de 4ºC por 20 minutos. A fração da solução foi transferida para um novo tubo e armazenada a -20ºC. Amostras de proteínas solúveis foram colocadas sobre gel NuPAGE® Bis-Tris 12% e processadas por digestão de tripsina no próprio gel[8] como preparação para o procedimento de análise por espectrometria de massa. A espectrometria de massa foi realizada no espectrômetro de massa híbrido *LTQ-FT-MS Ultra System*[9] e análise subsequente executada com o software de análise Scaffold[10]. No software Scaffold, as configurações para identificação das proteínas foram ajustadas da seguinte maneira: 99% para o *protein threshold*, *minimum number of peptides* igual a 2, e *peptide threshold* de 95%. Em análises subsequentes utilizou-se os valores quantitativos do espectro total normalizado (*normalized total spectra*). O programa Scaffold possui uma versão apenas para visualização de dados experimentais, a qual está disponível de forma gratuita para download a partir do website Proteome Software (Proteome Software, 2014). A análise de proteômica foi repetida duas vezes para cada amostra temporal e condição experimental. Valores médios foram utilizados nos cálculos de FBA. Os dados experimentais estão disponíveis, juntamente com os respectivos códigos de análise através da publicação (MONTEZANO et al., 2015).

Passamos agora a descrever o método proposto para determinação da função objetivo com dados experimentais de proteômica.

Assume-se inicialmente que há medições de expressão proteica

[4] *DMSO*: dimetilsulfóxido.

[5] Mefloquina obtida de Sigma Co., St. Louis, MO, USA.

[6] *m/v*: massa por volume, i.e. concentração de massa.

[7] *v/v*: volume por volume, i.e. concentração de volume.

[8] Promega ProteaseMAX, surfactante e melhorador da ação de digestão por tripsina.

[9] *Thermo Scientific*, Kalamazoo, MI, USA.

[10] Proteome Software, Corvallis, OR, USA.

resultantes de um experimento de proteômica para uma determinada condição experimental e instante de tempo, em que os níveis de K proteínas foram medidos (como apresentado acima). Essas medidas são armazenadas em um vetor $\boldsymbol{p}$:

$$\boldsymbol{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix} \tag{3.5}$$

em que cada elemento p_k corresponde ao valor que representa o nível da proteína k na amostra[11]. Conforme mencionado anteriormente, ainda que experimentos de proteômica gerem resultados muito confiáveis (diferentemente de experimentos com micromatrizes de DNA, cuja SNR é muito mais baixa), o número K de proteínas identificadas é apenas um subconjunto do proteoma completo da célula, uma dificuldade causada por dois fatores principais: (1) o método experimental identifica proteínas por comparação de fragmentos de peptídeos com uma base de dados, de forma que apenas proteínas previamente identificadas são incluídas nos resultados e (2) nem todas as proteínas são suscetíveis ao procedimento de digestão.

Adicionalmente, ao analisar o metabolismo com FBA em várias condições experimentais é possível que algumas proteínas não sejam detectadas em uma amostra específica (e.g. proteína A pode não estar presente na amostra do grupo de *controle 6 horas* após exposição ao antibiótico, porém pode estar presente na amostra de *tratamento* do mesmo instante amostral), e portanto o valor dessas proteínas em todas amostras são ajustadas para um nível zero. Esse procedimento não altera o funcionamento da técnica, porém dessa forma o vetor $\boldsymbol{p}$, independentemente do subconjunto de proteínas observado, sempre possui a mesma dimensão K e o conjunto de proteínas utilizada no algoritmo é sempre o mesmo, o que facilita a implementação do método. Outro fator importante de salientar é que um experimento de proteômica quantifica tanto enzimas quanto proteínas de sinalização e regulação, porém o vetor $\boldsymbol{p}$ é composto apenas de proteínas com atividade enzimática e que estão presentes no modelo metabólico *in silico*. Embora essa seleção deixe de incluir um bom número de importantes proteínas reguladoras, essa informação pode ser incorporada a outros métodos já publicados (COVERT; SCHILLING; PALSSON, 2001; CHANDRASEKA-

[11] O nível de cada proteína presente na amostra é medido utilizando a contagem espectral normalizada.

RAN; PRICE, 2010). É importante notar, entretanto, que a comparação que realizaremos a seguir *não é* dependente dessa informação, visto que nenhuma das técnicas comparadas utiliza informação proteica de regulação ou sinalização.

Após a construção do vetor $\boldsymbol{p}$, o segundo passo do método proposto é a normalização dos dados proteicos pelo valor máximo, para que os coeficientes estejam normalizados dentro do intervalo $[0, 1]$, como indicado em (COLIJN et al., 2009). O vetor de níveis relativos de proteína $\tilde{\boldsymbol{p}}$ é dado por:

$$\tilde{\boldsymbol{p}} = \boldsymbol{p}/\max\{\boldsymbol{p}\} \tag{3.6}$$

em que $\tilde{\boldsymbol{p}}$ também possui dimensão K. Os valores desse vetor normalizado serão usados nas expressões fornecidas no modelo metabólico GSMN-TB, o qual determina a ação enzimática resultante no caso de reações catalisadas por mais de uma enzima, como será explicado a seguir. É importante compreender que cada valor em $\tilde{p}_k$ está associado a apenas *uma* enzima.

Após o procedimento de normalização, identifica-se no modelo metabólico GSMN-TB quais as reações catalisadas pelas enzimas representadas em $\tilde{\boldsymbol{p}}$. Os valores de cada elemento $\tilde{p}_k$ é utilizado na determinação do valor do coeficiente c_i da nova função objetivo. No caso mais simples em que uma única enzima catalisa uma reação, ajusta-se o coeficiente c_i igual ao valor $\tilde{p}_k$ correspondente. Para todos os outros fluxos, i.e. aqueles que não estão presentes em $\tilde{\boldsymbol{p}}$, os respectivos coeficientes são ajustados para zero, visto que eles representam proteínas que não foram observadas no experimento.

Situações mais complexas surgem quando uma reação metabólica (i.e. um fluxo) não é catalisado por uma única enzima, mas pela ação combinada de um conjunto de enzimas. Em tais situações, a determinação dos coeficientes c_i da função objetivo depende da relação entre enzimas e reações metabólicas. Para isso é necessário compreender como enzimas e reações estão relacionadas no modelo metabólico, e como utilizamos essa informação na definição da nova função objetivo que está sendo proposta. No modelo GSMN-TB, para as reações metabólicas catalisadas por um conjunto de enzimas, a ação combinada dessas proteínas é dada por uma expressão booleana. Essas expressões foram pesquisadas e identificadas individualmente pelos autores através de pesquisa na literatura correspondente. Como um exemplo desse tipo de reação, temos a reação de síntese de mio-inositol, identificada pelo número *R022* no modelo metabólico. Para essa reação, as proteínas

catalíticas são associadas na seguinte expressão:

$$\text{Rv0046c} \vee (\text{Rv2612c} \wedge \text{Rv1822}) \tag{3.7}$$

em que os operadores $\vee$ e $\wedge$ representam os operadores booleanos `OR` e `AND` respectivamente. As proteínas nessa expressão são as duas transferases *Rv1822* e *Rv2612c* e a sintase *Rv0046c*, todas envolvidas na síntese do inositol.

Nestes casos, para determinação dos coeficientes c_i da função objetivo é necessário primeiramente substituir os dados de proteômica do vetor $\tilde{\boldsymbol{p}}$ nas posições correspondentes da expressão booleana (similar à Eq (3.7)). A operação matemática para o `OR` resulta no máximo dos dois termos, ao passo que a operação `AND` produz como resultado o mínimo dos dois termos, de maneira análoga à proposta utilizada em (COLIJN et al., 2009). Os valores resultantes para ação combinada das enzimas são utilizados como valores finais dos coeficientes c_i correspondentes a cada uma das reações. Notamos que o vetor $\boldsymbol{c}$ terá um número de elementos não-nulos em geral diferente de K (o número de proteínas identificadas no experimento de proteômica), a não ser que todas as reações sejam do caso peculiar e específico mencionado anteriormente, em que a reação metabólica é catalisada apenas por uma enzima distinta. Além desse fato, em geral K será maior que o número de elementos não-nulos de $\boldsymbol{c}$ (reações catalisadas pelas enzimas correspondentes). Embora uma situação rara, é possível que, em algumas situações, o número de coeficientes não-nulos de $\boldsymbol{c}$ na função objetivo seja maior que o número de proteínas identificadas em $\tilde{\boldsymbol{p}}$, por exemplo, quando uma mesma enzima (ou combinação de enzimas) seja responsável por catalisar diversas reações no modelo. Como pode-se ver pela Eq. (3.3), cada coeficiente c_i multiplica o correspondente fluxo metabólico v_i.

Na Figura 17, apresentamos um diagrama em blocos que corresponde à explicação dos parágrafos anteriores, o qual explica o procedimento de cálculo dos coeficientes da função objetivo a partir de dados experimentais de proteômica. Na figura, indica-se o vetor $\boldsymbol{p}$ de dimensão K contendo as medições de abundâncias enzimáticas resultantes de um experimento de proteômica. O vetor $\tilde{p}$ contém os dados normalizados. Para cada reação metabólica do modelo, indentificam-se as enzimas responsáveis por sua catálise e a maneira de ação combinada, caso a reação seja catalisada por um conjunto de enzimas. Na figura apresentamos como exemplo a reação metabólica *R022*, catalisada por três enzimas, como indicado na explicação do texto. O coeficiente da

função objetivo que corresponde à essa reação é calculado utilizando as funções máximo e mínimo. O valor obtido é utilizado como coeficiente do fluxo correspondente, i.e. v_{22} no exemplo. O procedimento deve ser repetido para cada um dos coeficientes do vetor $\boldsymbol{c}$, de forma a obter uma combinação linear de fluxos v_i que deve ser maximizada por FBA. Todos os coeficientes de reações sem representatividade no vetor $\boldsymbol{p}$ são colocados em zero e portanto não contribuem para o processo de otimização.

Embora simples, veremos que essa ideia produz melhorias em alguns aspectos importantes das predições do FBA. Para ilustrar o procedimento de determinação da função objetivo a partir de dados de proteômica, apresentamos a seguir um exemplo. Suponha que no experimento de proteômica sejam identificadas três proteínas 1, 2 e 3 com níveis $p_1 = 1.0$, $p_2 = 2.0$ e $p_3 = 3.0$ em uma determinada amostra. O vetor normalizado é obtido com a Eq. (3.6) e é igual a $\tilde{\boldsymbol{p}} = [1/3; 2/3; 1]$. Apenas para este exemplo, assumiremos que as três proteínas catalisam três reações do modelo metabólico GSMN-TB, a saber, *R023*, *R042* e *R128*. Assume-se ainda que a proteína 1 catalisa a reação v_{23}, a proteína 2 catalisa a reação v_{42} e uma combinação `AND` das proteínas 1 e 3 catalisa a reação v_{128}. De acordo com esses dados, a função objetivo a ser maximizada or FBA é definida como:

$$f(\boldsymbol{v}) = \frac{1}{3}v_{23} + \frac{2}{3}v_{42} + \frac{1}{3}v_{128} \tag{3.8}$$

em que o vetor de coeficientes é dado por $\boldsymbol{c} = [1/3; 2/3; 1/3]$. visto que a expressão `AND` é calculada utilizando o mínimo entre os dois valores em $\tilde{\boldsymbol{p}}$.

É interessante observar que, por exemplo, ao utilizar dados de proteômica amostrados em diferentes instantes de tempo, é possível definir funções objetivo (i.e. objetivos metabólicos) para diferentes momentos da vida da célula, algo que não é possível para a função objetivo de biomassa ou para qualquer outra função objetivo que seja independente do instante amostral.

Tendo apresentado a proposta de uma função objetivo definida por dados de proteômica, apresentaremos na próxima sessão os resultados de simulação obtidos para o metabolismo do Mtb utilizando esta metodologia.

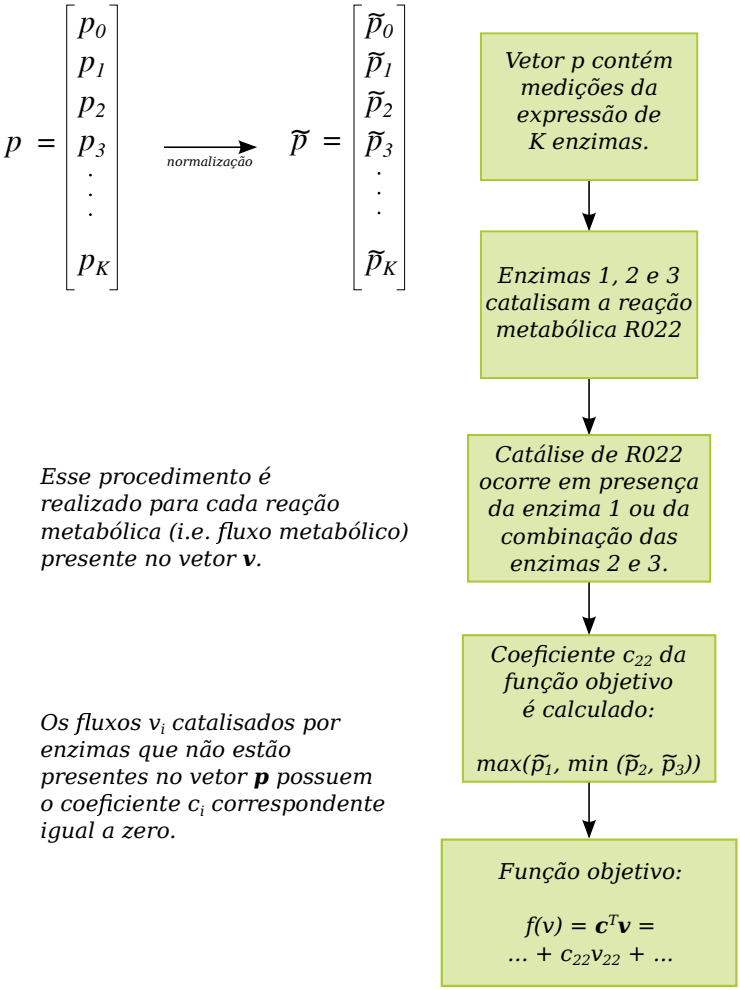

Figura 17 – Procedimento para determinação dos coeficientes da combinação linear de fluxos utilizada como função objetivo de FBA. A partir de medidas de abundâncias enzimáticas obtidas com um experimento de proteômica, os coeficientes são calculados a partir dos valores normalizados de acordo com o tipo de ação enzimática. Neste exemplo, o coeficiente para a reação metabólica *R022* é obtido da combinação dos valores de três enzimas (explicações no texto). *Fonte: Elaborada pelo autor.*

3.5 RESULTADOS

Para demonstrar a utilidade da proposta de função objetivo apresentada neste capítulo foram realizadas simulações e comparações da metodologia proposta com o método alternativo E-flux (COLIJN et al., 2009). Esse método foi originalmente proposto usando dados de transcriptômica para ajustar as restrições do problema de otimização do FBA com base em uma analogia de "capacidade e vazão"[12]. Visto que o método E-flux também pode ser usado com dados de proteômica, conforme indicado no próprio artigo, as simulações a seguir foram executadas tanto para o método proposto quanto para o método E-Flux com o mesmo conjunto de dados experimentais, e os resultados obtidos são comparados. Primeiramente os dois métodos foram simulados com dados de um grupo de controle, juntamente com o procedimento FBA normal com objetivo de maximização da biomassa e sem uso de dados experimentais. Neste trabalho assume-se que os dados de proteômica para o grupo de controle não devem ser significativamente diferentes para diferentes pontos amostrais (desconsideram-se as diferenças entre as fases de crescimento das bactérias), e todos os dados são considerados como realizações de uma mesma variável aleatória.

É importante ressaltar que nosso método foca em definir uma função objetivo a partir de dados de proteômica. Dessa forma, utilizamos as mesmas restrições de capacidade utilizadas no procedimento FBA básico, sem dados experimentais. Sendo assim, quando apresentamos os resultados das comparações entre o nosso método e o método alternativo E-flux, estamos comparando duas variantes do método FBA básico modificadas pela introdução de dados de proteômica: um método (E-flux) utiliza os dados experimentais para redefinir restrições de fluxo e uma função objetivo de biomassa, enquanto o método proposto utiliza dados de proteômica para definir a própria função objetivo e mantém fluxos intracelulares em restrições (como no procedimento FBA básico).

Para comparação, o valor de biomassa foi restringido ao mesmo valor obtido com o procedimento FBA básico. Os resultados das simulações foram analisados em termos do número de reações do modelo que apresentaram fluxo zero no vetor ótimo $\boldsymbol{v}^*$ obtido e são catalisadas por enzimas essenciais para o crescimento, conforme apresentado em (SASSETTI; BOYD; RUBIN, 2003). Após calcularmos a configuração metabólica ótima com FBA, identificam-se todos os fluxos que resultaram em valor zero (i.e. fluxos que não contribuem para a maximização da

[12] *"pipe capacity"* no original.

função objetivo). A partir do modelo metabólico sabemos quais enzimas catalisam as reações correpondentes a esses fluxos. Verificando em (SASSETTI; BOYD; RUBIN, 2003) se essas enzimas são essenciais para a sobrevivência do organismo, temos uma aparente contradição: uma reação metabólica que apresenta fluxo nulo, entretanto a presença da enzima correspondente é essencial à sobrevivência do organismo. Ainda que não seja possível concluir de forma defnitiva que a essencialidade da enzima esteja associada apenas com a catálise da reação, em muitas situações esse será o caso. Dessa forma, assume-se neste trabalho que tal seja o caso, para fins de comparação dos métodos.

Nestas simulações a análise está restrita à condição experimental do grupo de controle, em que as bactérias não foram expostas ao antibiótico e portanto assume-se a maximização da produção de biomassa como objetivo metabólico da célula. É esperado que em tal situação, quando o organismo não está sob a pressão de estresses biológicos, um número pequeno de *reações essenciais* apresentem fluxo zero. Por reações essenciais, assume-se reações metabólicas catalisadas por enzimas essenciais, i.e. enzimas que ao serem eliminadas do citoplasma (e.g. por algum processo de mutagênese) impedem a sobrevivência do organismo. Para um dos estudos atuais mais detalhados sobre essencialidade do Mtb, ver (SASSETTI; BOYD; RUBIN, 2003). Essencialidade de enzimas, tanto para situações *in vitro* quanto *in vivo*, podem também ser obtidas diretamente de bases de dados específicas do organismo, como em (TB Database, 2015b; Swiss Institute of Bioinformatics and École Polytechnique Fédérale de Lausanne, 2013).

Utilizando a rede metabólica *in silico* em escala genômica GSMN-TB do Mtb, uma simulação FBA foi executada para cada uma de três diferentes técnicas: (1) FBA básico com maximização da produção de biomassa, (2) método proposto com dados de proteômica determinando a função objetivo e (3) método E-flux com dados de proteômica definindo restrições de capacidade. As simulações (2) e (3) foram realizadas com dados do grupo de controle (em que crescimento *pode* ser assumido como o objetivo metabólico da célula). A função objetivo da simulação (2) foi definida com o método detalhado na seção anterior. Para a simulação (3) o método empregado para definir as restrições é exatamente o mesmo descrito em (COLIJN et al., 2009). Embora não seja o melhor caso de uso para o método proposto, como primeiro resultado é interessante avaliar se uma função objetivo definida com dados de proteômica ao invés do objetivo de biomassa ainda produz resultados que sejam biologicamente viáveis. Uma maneira é observar o número de reações com fluxo zero e que são catalisadas por enzimas essenciais ao cresci-

mento. Se assumirmos que as enzimas essenciais, em geral específicas para uma determinada reação metabólica, *devem* estar presentes no ambiente celular para que o crescimento seja garantido, não é provável que as reações correspondentes catalisadas por tais enzimas apresentem fluxo nulo, então quanto menor este número, mais próxima da realidade biológica será a distribuição obtida com FBA. Chamaremos estas reações metabólicas que são catalisadas por enzimas essenciais e apresentam fluxo nulo de ZFER (*zero-flux essential-enzyme reaction*), e utilizamos esse número como uma primeira métrica da qualidade dos resultados de FBA. Por trás da utilidade dessa métrica está a suposição de que reações catalisadas por enzimas essenciais não devem apresentar fluxo zero quando o objetivo metabólico é o crescimento bacterial. Na Tabela 3.5 apresentamos o número de ZFERs obtidos com cada técnica. Com o método FBA básico obtém-se 54% de reações do mo-

Tabela 5 – Número de reações catalisadas por enzimas essenciais conforme (SASSETTI; BOYD; RUBIN, 2003) e que apresentam fluxo igual a zero.

	FBA	PmxObj	E-flux
% ZFERs	54%	46%	57%
enzimas	100	66	103

▶ Esta tabela mostra a porcentagem das reações presentes no GSMN-TB catalisadas por enzimas essenciais e que apresentam fluxo nulo na distribuição de fluxos ótima obtida com cada um dos métodos FBA. O número de enzimas que catalisam essas reações é apresentado na segunda linha da tabela. Três métodos foram simulados: FBA básico com função objetivo de biomassa, método proposto com função objetivo definida com dados de proteômica (PmxObj) e método E-flux com restrições ajustadas por dados de proteômica.

delo com fluxo nulo. Para o método E-flux essa porcentagem sobe para 57%, embora esse aumento represente apenas a inclusão de mais 3 enzimas essenciais. O resultado mais interessante é que com o método proposto, o número de ZFERs diminui significativamente, para 46% do total de reações do modelo, representando um pouco mais que a metade do número de enzimas envolvidas nos outros dois métodos (66 enzimas essenciais). Embora as conclusões devam ser tomadas com cuidado, essa redução no número de ZFERs provavelmente aproxima as predições de FBA do metabolismo real do organismo, em que nenhuma reação realmente apresenta fluxo nulo. Nota-se que na Tabela 3.5, o número de enzimas essenciais não acompanha as variações da porcenta-

gem porque várias enzimas apresentam ação combinada sobre algumas reações, como discutido na seção anterior em relação às expressões booleanas. Embora a proporção de ZFERs seja bastante significativa, essa métrica não é avaliada em artigos que estudam o metabolismo com FBA (e.g. (SCHUETZ; KUEPFER; SAUER, 2007)). O menor número de reações resultantes com fluxo zero quando utiliza-se a função objetivo de proteômica possivelmente é um indicativo da informação extra trazida para dentro do problema de otimização pelos dados. Nota-se que de todas as reações que ainda produzem fluxo zero, mesmo com a nova função objetivo, a maior parte foi também observada com o objetivo de biomassa no método FBA básico, mostrando que a função proposta não induz reações *diferentes* a zero, mas apenas reduz o número destas quando comparado com a maximização de biomassa, trazendo assim, mais informação biológica para o problema.

Um outro aspecto a analisar é o papel da função objetivo de biomassa (cf. Eq. (3.1)). No GSMN-TB, essa é uma reação de transporte definida como uma reação de conveniência que agrega proporções de vários substratos. Quando deixamos de utilizar a função objetivo de biomassa para uma função definida com dados de proteômica (que afeta principalmente as reações intracelulares de metabólitos precursores dos componentes da Eq. (3.1)), o consumo e a produção desses componentes da biomassa não serão mais maximizados. Como esses metabólitos não estão mais sob a necessidade de serem maximizados eles não estarão sob demanda, e é possível que as reações intermediárias que os sintetizam tenham os seus fluxos reduzidos ou mesmo levados a zero. Por esse motivo, alguns fluxos do vetor ótimo $\boldsymbol{v}^*$ serão forçosamente diferentes do que quando maximizamos biomassa. Visto que no caso de definir a função objetivo com dados de proteômica de controle o fenótipo resultante pode mesmo apresentar fluxo zero para a biomassa, é importante restringir esse fluxo para um determinado valor mínimo para que as conclusões entre os métodos possam ser comparadas. Sendo assim, nas simulações (2) e (3) o fluxo de biomassa foi restringido ao mesmo valor obtido com o FBA básico. Em geral, essa restrição mínima altera principalmente fluxos de conveniência associados com a produção de biomassa. Visto que essas reações de conveniência alteram alguns fluxos de biossíntese que estão acima na cadeia de reações, no caso de estudarmos condições experimentais de controle com a função objetivo de proteômica, é importante, para realizar a comparação, que estes fluxos de biossíntese estejam sob demandas similares. A necessidade dessa restrição justifica-se para realizarmos de forma adequada a comparação entre métodos.

Dito isso, reiteramos que o objetivo ao sugerir uma função definida com proteômica *não é* substituir a maximização de biomassa quando esta é um objetivo metabólico adequado, mas sim em situações de sobrevivência. Como uma fonte de validação entretanto, seria razoável esperar que, ao utilizarmos dados de um grupo de controle para definir a função objetivo, o funcionamento metabólico da célula se mantenha.

Visto que os dados de proteômica estão disponíveis para diferentes instantes amostrais, como um novo passo de validação do método proposto, foram realizadas comparações do erro de predição entre o nosso método e o método E-flux, para diversas condições experimentais de controle e tratamento. As simulações de FBA para os dois métodos foram feitas tomando-se subconjuntos dos dados de proteômica selecionados aleatoriamente e usando as proteínas restantes para avaliar o erro de predição do modelo em termos de abundância enzimática. Esse procedimento de validação cruzada foi realizado da forma explicada a seguir. Inicialmente, para um determinado instante amostral e condição experimental (controle ou tratamento), o conjunto de proteínas do vetor $\boldsymbol{p}$ (aproximadamente 400 medições para este experimento) foi dividido de forma aleatória em dois grupos, treinamento e validação, usando uma regra 80/20. O grupo de treinamento, vetor $\boldsymbol{p}_t$, possui dimensão K_t e contém 80% dos valores das proteínas em $\boldsymbol{p}$. O grupo de validação, vetor $\boldsymbol{p}_v$, possui dimensão K_v e contém os 20% restantes dos valores de $\boldsymbol{p}$.

O vetor de treinamento $\boldsymbol{p}_t$ foi utilizado para definir a função objetivo do método proposto e para definir as restrições de capacidade do método E-flux. Realizou-se a seguir o procedimento de FBA para os dois métodos, obtendo-se um vetor de fluxos ótimo com o método proposto, $\boldsymbol{v}^*_{pmx}$ e um vetor de fluxos ótimo com o método E-flux, $\boldsymbol{v}^*_{efl}$.

Com as proteínas do grupo de validação $\boldsymbol{p}_v$, foram avaliadas as expressões booleanas correspondentes e os valores resultantes armazenados em um vetor de abundâncias enzimáticas para validação definido como $\boldsymbol{v}_a$. Os valores desse vetor foram utilizados para comparação com os fluxos correspondentes obtidos com FBA para cada método. A comparação, portanto, é feita entre fluxos metabólicos e abundâncias enzimáticas, visto não estarem disponíveis medidas de fluxos metabólicos.

O vetor erro de predição é calculado como a diferença entre o vetor $\boldsymbol{v}_a$ e os vetores resultantes de FBA para cada método, $\boldsymbol{v}^*_{pmx}$ e $\boldsymbol{v}^*_{efl}$. Calcula-se a norma desse vetor diferença e esta é o erro quadrático de predição.

Este procedimento é repetido 16 vezes com diferentes divisões das

proteínas de $\boldsymbol{p}$, i.e. diferentes grupos de treinamento $\boldsymbol{p}_t$ e validação $\boldsymbol{p}_v$. O erro quadrático médio de predição é calculado como a média dos erros das 16 realizações da simulação. O erro quadrático normalizado foi utilizado para avaliar a qualidade da predição dos dois métodos. O erro foi normalizado pelo número de reações no conjunto de validação, visto que sem a normalização, diferentes divisões dos dados resultariam em um diferente número de reações dependendo das expressões booleanas que seriam avaliadas.

Os resultados para o erro quadrático médio de predição (*mean square error of prediction* - MSEP) estão apresentados na Tabela 3.5. Estes são resultados médios de 16 realizações da simulação. Observa-se que o MSEP para o grupo de validação é menor para o método proposto em todas as condições experimentais e pontos amostrais. Valores de *p-value* do teste-*t* (nível de significância 95% e 15 graus de liberdade) para a diferença do erro são mostrados na última linha da tabela. Observou-se que mesmo para *p*-values mais altos, o erro é menor de forma consistente para a metodologia proposta, embora a diferença possa ser considerada pouco significativa. Mostramos na Fig. 18 um diagrama com a distribuição do erro obtido nas 16 realizações da simulação para os dois métodos.

Tabela 6 – Comparação do MSEP para o método proposto e o método E-flux.

	CTL	H6T	D2T	D4T
MSEP PmxObj	0.20	0.26	0.21	0.23
MSEP E-flux	0.24	0.34	0.24	0.26
p-value	0.07	0.02	0.13	0.20

▶ Esta tabela mostra o erro quadrático médio de predição (*mean square error of prediction* - MSEP) para o método proposto (PmxObj) e para o método E-flux com dados de proteômica para diferentes condições experimentais. O método proposto produz resultados com menor erro de predição em todas as condições. (t-test com nível de significância de 95%, 15 graus de liberdade). **CTL** (controle), **H6T, D2T, D4T** (tratamento 6 horas, 2 dias e 4 dias após exposição à mefloquina. A última linha da tabela mostra *p*-values do teste "t"para as diferenças entre os erros.

Vê-se pela figura que embora haja sobreposição entre os resultados, em geral observam-se resultados médios com menor erro para a metodologia proposta. Finalmente, para esse conjunto de dados e número de realizações, ainda que para algumas divisões do conjunto de

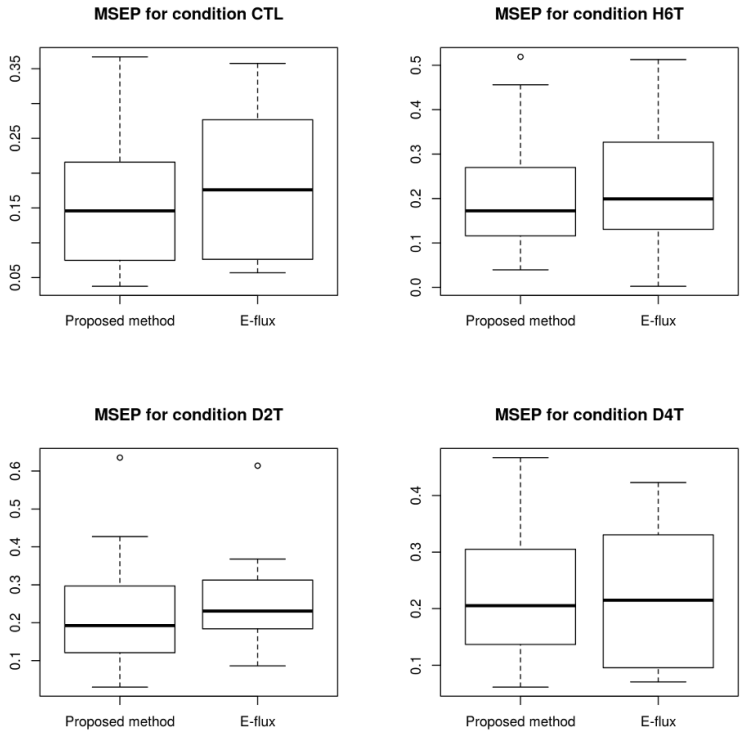

Figura 18 – Erro quadrático médio de predição para o método proposto e o método E-flux em três condições experimentais. O uso de dados de proteômica para definir a função objetivo em FBA resulta em menores erros de predição. *Fonte: Elaborada pelo autor.*

proteômica o erro para a metodologia proposta possa ser mais elevado que para alguma instância do método E-flux, em termos médios o uso da função objetivo de proteômica é melhor.

Em otimização baseada em restrições (*constraint-based optimization*) é possível que soluções ótimas não sejam únicas. No caso de FBA, ótimos alternativos representam situações em que o sistema metabólico atinge o mesmo valor para a função objetivo usando um conjunto diverso de reações, o que é possível devido às redundâncias inerentes à rede metabólica. Sendo assim, apresentamos a seguir o uso de análise de variabilidade de fluxos (*flux variability analysis* - FVA) (MAHADEVAN; SCHILLING, 2003) com a metodologia proposta e com o método E-flux para analisar o impacto de possíveis ótimos alternativos. Embora existam outras possibilidades de análise, nestas simulações decidiu-se avaliar a presença e o impacto de soluções ótimas alternativas com a técnica de FVA, a qual tem sido utilizada com sucesso para inferir sobre ótimos alternativos e está disponível para ser utilizada com o pacote SurreyFBA (GEVORGYAN et al., 2011) utilizado em todas simulações de FBA neste trabalho. Por exemplo, outra possibilidade de análise é o uso do método MILP que enumera todas as possíveis soluções alternativas ótimas. Essa técnicas entretanto, no caso de redes metabólicas em escala genômica como a deste trabalho, pode ser computacionalmente inviável devido ao aumento exponencial do número de pontos extremos (i.e. soluções) que podem existir (MAHADEVAN; SCHILLING, 2003; REED; PALSSON, 2004). FVA basicamente determina a faixa de variabilidade para cada fluxo da rede devido à presença de soluções ótimas alternativas, permitindo o estudo de importantes características do comportamento do sistema devido à redundância da rede.

Ótimos alternativos são responsáveis pela situação em que diferentes maneiras de utilizar a rede de reações metabólicas subjacente correspondem a uma mesma função celular (REED; PALSSON, 2004). Ao definirmos o problema de utilização como a maximização de um subconjunto de fluxos intracelulares ao invés de biomassa, espera-se que a variabilidade nos fluxos seja reduzida. Visto que a função de biomassa é ela própria uma combinação linear de diversos fluxos intracelulares (cf. Eq (3.1)), uma hipótese possível é que mais redundância esteja presente na maximização da biomassa, em contraste com a maximização de um conjunto de fluxos intracelulares propriamente dito. Esse resultado é o que observamos nas simulações de FVA. Na Tabela 7 mostramos que o uso de proteômica para definir a função objetivo de fato resulta em um menor número de fluxos com alta variabilidade. Definimos como reações com alta variabilidade de fluxo, todas aquelas reações para as

quais a faixa de variação do fluxo (i.e. fluxo máximo menos o valor de fluxo mínimo produzidos por FVA) resultante é maior que 0.05.

Tabela 7 – Número de reações com alta variabilidade de fluxo para o método proposto, com função objetivo definida por proteômica, e para o método E-flux.

Condição	E-flux	PmxObj
6 hrs CTL (repetição 1)	164	72
6 hrs CTL (repetição 2)	185	79
6 hrs MEF (repetição 1)	276	72
6 hrs MEF (repetição 2)	248	65
Dia 2 MEF (repetição 1)	149	72
Dia 2 CTL (repetição 2)	272	70
Dia 2 MEF (repetição 1)	266	70
Dia 2 MEF (repetição 2)	267	72
Dia 4 CTL (repetição 1)	259	72
Dia 4 CTL (repetição 2)	266	72
Dia 4 MEF (repetição 1)	276	68
Dia 4 MEF (repetição 2)	298	70

► Esta tabela mostra o número de fluxos metabólicos com alta variabilidade devido a soluções ótimas alternativas em FBA. A primeira coluna identifica a condição experimental e o ponto amostral, a segunda apresenta a quantidade de reações para o método E-flux, e a última coluna mostra os resultados para o método proposto neste capítulo. Com a metodologia proposta, o número de reações metabólicas que apresentam alta variabilidade é significativamente reduzido em comparação com o método E-flux para todas as condições. O número de reações para o método FBA básico (não mostrado) é similar aos números apresentados para o método E-flux, indicando que é a definição da função objetivo, e não as restrições, o que permite a redução do impacto das soluções ótimas. Os valores apresentados representam fluxos para os quais a faixa de variação (fluxo máximo menos fluxo mínimo) devido aos ótimos alternativos é maior que 0.05.

É interessante notar nesses resultados que o uso de função objetivo que maximiza fluxos específicos conforme as medidas observadas de proteômica reduz o impacto da redundância da rede na solução ótima quando comparado à função objetivo de biomassa. Observa-se que, com a função objetivo proposta, o objetivo da célula é menos disperso, visto que um número significativamente menor de reações permite alta

variabilidade no fluxo correspondente.

Além do número de reações, foi analisada também a variabilidade do fluxo nas reações catalisadas por enzimas essenciais, de acordo com o critério de essencialidade descrito em (SASSETTI; BOYD; RUBIN, 2003). Grande parte dessas reações catalisadas por enzimas consideradas essenciais apresentaram variabilidade zero nos fluxos correspondentes ou redução na magnitude da variabilidade em todas as simulações ao ser utilizada a função objetivo definida por dados de proteômica. Tais resultados indicam uma possível vantagem com a incorporação de dados de proteômica na função objetivo ao invés de sua utilização na definição de restrições, como propõe o método E-flux. Devido à menor variabilidade obtida na representação, o método apresenta vantagens quando é necessário discernir distribuições de fluxo metabólico biologicamente relevantes.

As simulações de FVA revelam que, além da redução no número de reações que apresentam alta variabilidade, o uso da função objetivo proposta também ajuda a reduzir as magnitudes dessas variações quando comparadas com o método E-flux. Na Fig. 19, mostramos o logaritmo dos valores médios de variabilidade em todas as condições experimentais para a metodologia proposta e para o método E-flux. Desses resultados confirma-se uma redução na variabilidade geral da distribuição de fluxos para o método proposto, produzindo uma solução ótima que é menos afetada pela presença de soluções ótimas alternativas. Em todas as condições as faixas de variabilidade para o método E-flux são maiores que com a técnica proposta.

Os resultados obtidos, ainda que não definitivos, indicam que há vantagens no uso de funções objetivo definidas por dados experimentais de proteômica, principalmente em situações de sobrevivência celular (e.g. exposição a antibióticos), quando maximização de biomassa não é o melhor objetivo metabólico que pode ser assumido. Acreditamos que mais pesquisas nessa direção são úteis e necessárias para melhor avaliar essa nova e promissora proposta no FBA.

3.6 DIFICULDADES DO MÉTODO

Da mesma forma como em outros métodos que não utilizam informações de regulação genética em FBA, o uso da função objetivo de proteômica pode apresentar resultados espúrios no caso de proteínas que possuem atividade tanto metabólica quanto regulatória ou de sinalização. Algumas enzimas, como a proteína *katG – Rv1908c*, apre-

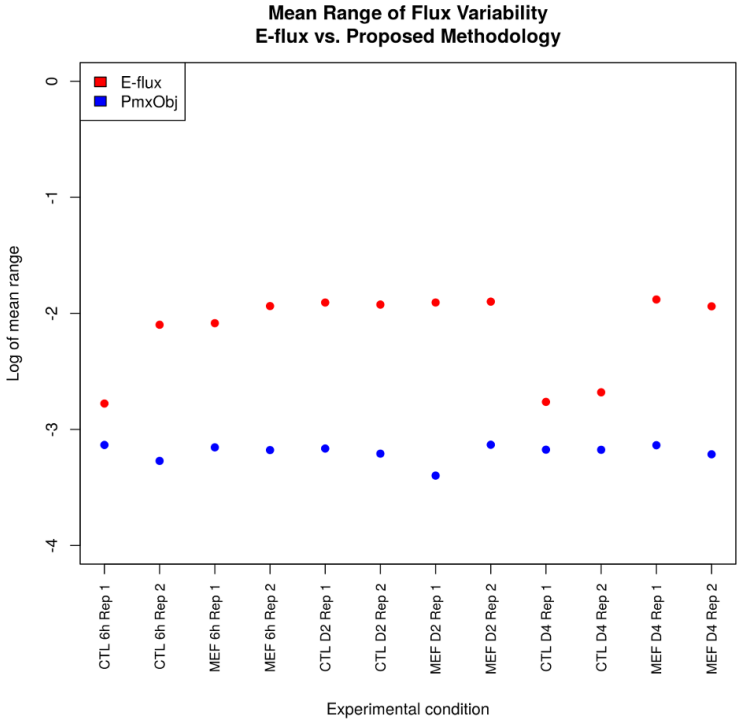

Figura 19 – Comparação da variabilidade de fluxo entre o método proposto e o método E-flux para todas condições experimentais. O uso de dados experimentais de proteômica na definição da função objetivo de FBA resulta em menor variabilidade média devido a soluções ótimas alternativas. *Fonte: Elaborada pelo autor.*

sentaram neste conjunto de dados altos valores de expressão proteica, e quando utilizadas para construir a função objetivo com dados de proteômica, aumentam significativamente as correspondentes reações metabólicas catalisadas. Entretanto, é possível que o motivo da alta expressão não seja suprir atividade enzimática, mas antes exercer um papel regulatório necessário durante a reprogramação metabólica. A proteína *katG* apresentou altos níveis de expressão proteica na condição de tratamento com a bactéria exposta à mefloquina, aumentando significativamente os fluxos das reações R134 e R363. Entretanto, os altos valores de fluxo de outras reações em caminhos similares não corroboram estes altos valores de fluxo, sendo possível que os altos níveis dessa proteínas estejam mais ligados a uma necessidade regulatória do que catalítica. Os motivos pelos quais um alto nível de expressão proteica não se traduz de forma inequívoca em maiores fluxos metabólicos são diversos, mas isso pode ocorrer devido, por exemplo, à presença de moléculas inibidoras de atividade enzimática (impedindo que a enzima entre em sua conformação ativa), ou à presença ou ausência de outras proteínas anteriores na cadeia regulatória. Embora o uso de dados experimentais em FBA já tenha sido bastante estudado, a incorporação de informações regulatórias mais precisas, de forma automática e com abrangência em escala genômica, ainda deve ser um foco ativo de pesquisa.

Duas outras possíveis causas de interferência e degradação da qualidade dos resultados de FBA são: (1) o limitado número de proteínas que um experimento de proteômica pode identificar em comparação com o proteoma completo da célula e (2) o próprio modelo *in silico* (GSMN-TB) utilizado como base para as simulações, o qual é incompleto, com partes da rede metabólica que estão ausentes por não terem sido ainda bem caracterizadas ou por não terem sido implementadas no modelo, ou ainda porque muitas reações metabólicas não possuem enzimas específicas associadas. Finalmente, pode haver inconsistência nas anotações presentes nas bases de dados.

Em geral, as dificuldades enfrentadas no uso de FBA para estudo de modelos metabólicos em escala genômica reside no fato que a quantidade de informação necessária no modelo é grande e na maioria dos casos necessita de validação experimental individual. O processo, desde a identificação das reações metabólicas, enzimas associadas, anotação, validação e criação do modelo exige uma siginificativa quantidade de processamento manual, especialmente lento e sujeito a erros. Tais problemas aumentam significativamente no caso da incorporação de dados experimentais e informações regulatórias e de sinalização. FBA

pode ser utilizado para estudo de caminhos parciais como apresentado em (VARMA; BOESCH; PALSSON, 1993; COVERT; SCHILLING; PALSSON, 2001; COLIJN et al., 2009), e é possível que em tais casos a utilidade da técnica seja melhor apreciada, visto que em tais casos as informações disponíveis são mais detalhadas e completas. Apesar das dificuldades associadas com esses modelos, o seu uso tem produzido inferências úteis sobre os organismos. Assim, deve-se buscar uma melhoria constante e iterativa dos modelos e das técnicas experimentais e computacionais. Essas poderão, com o tempo, levar à obtenção de resultados mais precisos e abrangentes a partir dos modelos metabólicos em escala genômica (PALSSON, 2000).

3.7 CONCLUSÃO

Neste capítulo foi avaliado o uso de dados de proteômica para determinar uma função objetivo para o método de análise de balanço de fluxos que seja dependente da condição experimental e, portanto, específica para cada situação metabólica. Algumas vantagens dessa proposta são:

- ao definir uma função objetivo com base em dados de proteômica obtém-se estimativas de valores de fluxo que representam diretamente variações no metabolismo causadas por variações no conteúdo proteico. Este é um importante aspecto do método para o caso de análise de fenótipos diferenciais, quando deseja-se uma comparação das diferenças em metabolismo entre um grupo de controle e um grupo de tratamento;

- a análise do metabolismo em diferentes instantes amostrais é facilitada quando utiliza-se medidas de proteoma no FBA, pois estas são intimamente relacionadas com os fluxos no caso de organismos procariotas (descartando-se modificações pós-traducionais e modulação de metabólitos). A técnica proposta permite definir uma função objetivo para cada condição experimental, e obter um fenótipo ótimo do regime permanente da célula para momentos específicos da vida celular;

- a técnica proposta também pode ser combinada com outras estratégias. Por exemplo, um termo adicional de produção de biomassa pode ser adicionado ao vetor $\boldsymbol{c}$, simplesmente alterando o valor do coeficiente c_{bio}. Esse termo pode ser usado para especificar uma pequena contribuição de biomassa ao objetivo, quando

a sua produção não é o objetivo metabólico principal do organismo. Produção de biomassa, ou qualquer outro fluxo, também pode ter seus limites máximo e mínimo ajustados nas restrições do problema de otimização, como utilizado no método E-flux.

Comparações foram realizadas em termos do número de reações e enzimas essenciais para o crescimento (conforme publicação de Sassetti et al. (SASSETTI; BOYD; RUBIN, 2003)), para dados de proteômica dos grupos de controle e em termos do erro de predição com o método E-flux (COLIJN et al., 2009), uma técnica alternativa que utiliza dados de proteômica para ajuste das restrições de FBA. Observou-se que ao utilizar uma função objetivo definida com dados de proteômica, FBA produz distribuições de fluxo com um menor número de reações com fluxo zero catalisadas por enzimas essenciais para o crescimento. Essa função objetivo leva também de forma consistente a um menor erro de predição em comparação com o método E-flux. Utilizando a técnica de FVA (*flux variability analysis*), observou-se que com a função objetivo proposta é possível reduzir a variabilidade dos fluxos e o impacto de ótimos alternativos na solução de fluxos $\boldsymbol{v}$. Esse resultado é importante, visto ser desejável que a incorporação de dados experimentais ajude a reduzir a incerteza na identificação das distribuições metabólicas relevantes em diferentes condições experimentais. Os resultados de FVA também mostram que com a função objetivo proposta obtemos um menor número de fluxos com alta variabilidade bem como variabilidade com reduzida magnitude para toda a distribuição de fluxos.

Embora neste estudo as simulações de FBA tenham sido feitas apenas com proteínas com atividade enzimática, conforme mencionado, é provável que a inclusão de proteínas de sinalização e regulação contribua para diminuir as dificuldades associadas com proteínas que possuem papéis tanto enzimáticos quanto de sinalização dentro da célula.

Ainda que a técnica proposta não possa ser considerada como a solução final para a determinação de função objetivo em FBA, o estudo realizado neste capítulo serve para colocar em discussão os limites da função de biomassa para casos quando crescimento ótimo não pode ser assumido como objetivo metabólico da célula, e a possibilidade de usar dados experimentais para guiar esse objetivo metabólico. Os resultados obtidos indicam que há claras vantagens no uso da função objetivo proposta, principalmente em casos de estresse biológico, como exposição a antibióticos. Acreditamos que mais pesquisas nessa direção são necessárias para que haja uma melhor avaliação dessa nova e promissora técnica.

4 DESCRITORES METABÓLICOS

4.1 INTRODUÇÃO

Uma solução possível para aprimorar a qualidade das inferências a partir de dados com formato *p-grande,n-pequeno* é a redução da dimensão p do problema. Essa redução da dimensionalidade de um conjunto de dados multivariáveis pode ser obtida através de uma de duas vias possíveis: (1) mantendo o mesmo tipo de dados e selecionar apenas um subconjunto das variáveis, ou (2) alterar o tipo de dados determinando um novo conjunto de descritores de menor dimensionalidade a partir dos dados originais.

Um exemplo da primeira técnica é o truncamento da dimensão para seleção de um conjunto menor de variáveis. Assumindo um vetor observado de dados $\boldsymbol{x}_i = [x_1 x_2 \dots x_D]$, sendo D a dimensão do vetor e cada componente $x_k \in [0,1] \forall k \in 1, 2, \dots, D$, podemos selecionar um vetor de menor dimensão que contenha apenas os d maiores valores de x_k, em que $d < D$. É possível também selecionar utilizando-se um valor de corte (*threshold*) λ e descartar todas as componentes para as quais $x_k < \lambda$.

Como exemplo da segunda via de redução de dimensionalidade, podemos citar o uso de *análise de componentes principais* (principal component analysis - PCA), em que os dados são linearmente transformados e projetados sobre direções de menor dimensionalidade e que mantêm a maior quantidade de informação possível. Diversas técnicas de mapeamento existem ainda na forma de ortogonalizações e também com o uso de kernels (THEODORIDIS; KOUTROUMBAS, 2008; SCHÖLKOPF; SMOLA, 2002).

Neste capítulo iremos apresentar uma técnica do segundo tipo, em que um conjunto de dados de alta dimensionalidade é utilizado para determinar descritores não apenas de menor dimensionalidade, mas que também representam o metabolismo do organismo em um nível mais alto que uma distribuição de fluxos determinística, como foi o foco do capítulo anterior. Esses novos descritores são probabilísticos e representarão a atividade molecular de todo um caminho metabólico.

A utilidade de descritores deste tipo é dupla. Além da desejada redução na dimensionalidade, a informação sobre a atividade de um caminho metabólico ao invés de fluxos metabólicos individuais, como veremos, é de mais fácil interpretação e permite uma melhor compreensão da fisiologia do organismo e das alterações metabólicas causadas

por mudanças nas condições do ambiente dentro do qual o organismo está inserido, como disponibilidade de nutrientes, mudanças de temperatura, pressão e pH, exposição a compostos tóxicos e compostos bactericidas, entre outros.

O uso de descritores de alto nível para caminhos de sinalização celular foi apresentado em (EFRONI et al., 2011), e forma a base para o método que apresentamos neste capítulo, em que combinamos a ideia de probabilidade de ativação de um gene com a regulação em cascata das redes de regulação genética e metabólica.

É um novo método que utiliza conhecimento a priori sobre a estrutura de caminhos metabólicos para análise do metabolismo do *Mycobacterium tuberculosis*. O método é baseado em outro método previamente publicado no contexto de sinalização celular em estudos de câncer. Um importante uso do método apresentado é a descoberta de caminhos de sobrevivência do *Mycobacterium tuberculosis* quando exposto a drogas antibióticas, levando a uma possível solução para o desenvolvimento de novos compostos bactericidas de ação sinergística que podem auxiliar no tratamento da tuberculose.

Conhecimento do metabolismo celular é chave para compreender o papel da atividade de cada organismo no ambiente. Em particular, o conhecimento do metabolismo dos procariotas, bem como da sua fisiologia, são essenciais para elucidar suas relações patogênicas e parasitárias com outros organismos. Apesar do tamanho e da simplicidade e destes organismos unicelulares, eles representam uma enorme diversidade e plasticidade metabólicas, permitindo que muitos organismos, tal como o *Mycobacterium tuberculosis*, tenham uma alta capacidade de adaptação e sobrevivência em ambientes diversos (KIM; GADD, 2008). Embora muitas drogas estejam disponíveis para o tratamento da tuberculose e vários outros candidatos tenham sido aprovados nos seus primeiros testes clínicos (OSBORNE, 2013), em muitos casos o bacilo causador da doença frequentemente torna ineficazes essas drogas devido à sua alta adaptabilidade. Como discutido no capítulo de Introdução, o desenvolvimento das cepas resistentes está intimamente ligado ao fato de pacientes interromperem o tratamento antes da cura completa (CASTELNUOVO, 2010). Visto que o tratamento completo da tuberculose dura geralmente vários meses, um ponto-chave do problema que necessita ser solucionado é o encurtamento da duração dos tratamentos.

A forma mais direta de lidar com este problema é o desenvolvimento de novos compostos antibióticos com ação sinergística, os quais consigam manter a taxa de depleção das bactérias mesmo após o início

do desenvolvimento da resistência às drogas de linha de frente geralmente prescritas, como mostra o diagrama da Figura 20. Assume-se que as bactérias ativam alguns caminhos metabólicos e inibem a ação de outros na tentativa de maximizar suas chances de sobrevivência. Os caminhos metabólicos que são ativados pelo estresse (no caso da figura, exposição a um antibiótico A), são prováveis caminhos com funções necessárias à sobrevivência, aos quais denominamos provisoriamente caminhos de *escape*, ou de *sobrevivência*. É possível que a observação cuidadosa dos caminhos que se mantêm ativados nas colônias resistentes forneçam bons alvos para o desenvolvimento de novos compostos que irão bloquear a atividade nestes caminhos de escape de uma maneira sinergística, produzindo como efeito uma depleção mais rápida das bactérias e menor duração no tratamento para a tuberculose. Na Figura 20, os círculos maiores representam uma célula (i.e. uma bactéria). A célula na condição **A** (esq.) representa uma célula sob condições normais de crescimento. A célula na condição **B** (ao centro) representa uma célula sob a ação bactericida de um primeiro composto antibiótico, enquanto a célula na condição **C** representa uma célula em processo de morte devido à exposição a uma segunda droga de ação sinergística. Os círculos menores dentro de cada célula representam caminhos metabólicos. Por exemplo, o caminho 1 pode representar o conjunto de reações metabólicas que produzem energia na célula, o caminho 2 pode conter as reações que trabalham para a biossíntese de biotina, ao passo que o caminho 3 poderia representar o conjunto das reações necessárias para a degradação de mRNA ou produção de um determinado amino-ácido, como valina, e assim por diante para todos os caminhos. Assume-se que na condição **A** (esq.), quando o objetivo principal da célula é o crescimento e a reprodução e nenhum estresse está presente, todos os caminhos metabólicos apresentam níveis normais de atividade, e são representados pela cor azul. Quando a bactéria é exposta a um primeiro antibiótico A, tal situação leva a célula para um novo estado no qual o seu objetivo principal é alterado para a sobrevivência, como ocorre com qualquer organismo. Esse novo fenótipo apresenta à célula a necessidade de encontrar uma nova configuração metabólica que maximize as suas chances de sobrevivência, não mais o crescimento ótimo e a reprodução. Para atingir esse objetivo, a atividade metabólica é inibida (e.g. para economizar recursos celulares ou talvez porque tal caminho ative uma pró-droga), enquanto a atividade em outros caminhos é aumentada (e.g. reações metabólicas que reduzem a atividade bactericida do antibiótico, ou produzem atividades de 'limpeza', como detoxificação e efluxo que secretam o antibiótico para

fora da célula, não permitindo a sua acumulação). Na figura, caminhos com a atividade aumentada são indicados na cor verde, e caminhos com inibição de atividade metabólica são representados em cinza. Sabe-se que o *Mycobacterium tuberculosis* é um patógeno altamente adaptável, e que possui a habilidade de ajustar o seu metabolismo para sobreviver sob muitos diferentes tipos de situações biologicamente estressantes, que como já discutido, é uma das razões pelas quais os tratamentos para a tuberculose são tão longos, causando interrupção do tratamento, re-infecção e finalmente o desenvolvimento de linhagens resistentes ao antibiótico. Nessa situação é que uma combinação de drogas geralmente é necessária para a obtenção da cura completa. Notando os caminhos que apresentam atividade aumentada devido à exposição ao primeiro composto, e que são responsáveis pela sobrevivência celular, é possível desenvolver um novo composto que atue sobre determinadas enzimas desses caminhos com o objetivo de bloquear a sua atividade, produzindo uma depleção maior e mais rápida das colônias e eventualmente a morte, como mostra a condição **C** (dir.)).

A ação específica de um composto bactericida em geral causa a depleção de um número de organismos, porém uma parcela das bactérias, em alguns casos, pode desenvolver resistência ao composto através da modificação do seu metabolismo, utilizando vias metabólicas alternativas que não são comprometidas pela ação do composto. A utilização de um composto de ação sinergística, que atue sobre o fenótipo de sobrevivência sem modificar a ação do composto original é extremamente desejável, pois permitirá a continuidade da ação bactericida e um tratamento mais curto e eficaz, diminuindo por sua vez a possibilidade de desenvolvimento de novos mecanismos de resistência.

A reinfecção e o aparecimento de cepas resistentes estão associados com a capacidade de adaptação da bactéria e as subsequentes alterações no metabolismo (precedidos por uma cascata de alterações em regulação, níveis de transcrição e de expressão proteica) na presença de estresses ambientais. São abundantes os estudos e as pesquisas para compreender os mecanismos de ação de possíveis compostos anti-tuberculose candidatos quanto para compreender os mecanismos de resistência desenvolvidos pelo Mtb a esses mesmos compostos e a outros antibióticos de primeira linha já aprovados, como por exemplo, a isoniazida (SLAYDEN; BARRY, 2000; TIMMINS; DERETIC, 2006; VILCHIÈZE; JACOBS, 2007). Dessa forma, compreender as alterações metabólicas que produzem os fenótipos resistentes é um passo crucial para combater com sucesso a tuberculose, tão importante quanto compreender o mecanismo de ação das drogas candidatas e já aprovadas

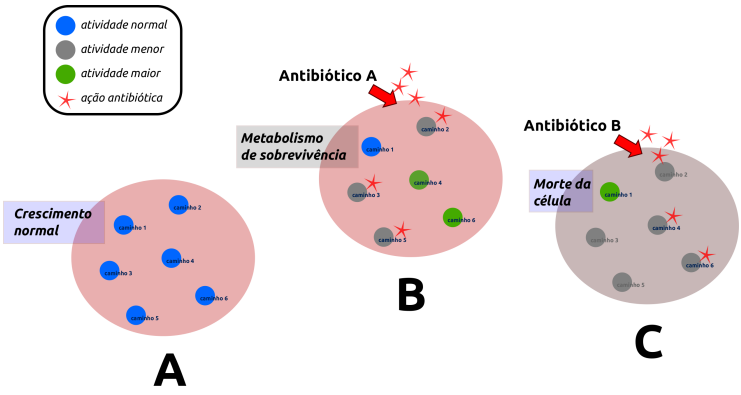

Figura 20 – Diferentes configurações metabólicas são colocadas em funcionamento pela célula para lidar com diferentes condições ambientais, tais como exposição a antibióticos. Caminhos metabólicos que sofrem alterações significativas de uma condição de controle para uma condição de estresse biológico podem ser a chave para identificação de novos alvos metabólicos para drogas de ação sinergística. *Fonte: Elaborada pelo autor.*

(GOTTLIEB; SHAW, 1967).

Todavia, para compreender o metabolismo no seu todo e desvendar a sua complexidade, não basta apontar genes individuais ou proteínas que variem a sua expressão de uma condição para outra. Técnicas de escopo mais amplo são necessárias, as quais tentem obter e produzir informações de mais alto-nível na hierarquia do comportamento celular. É nessa direção que apresentamos neste capítulo um método para estudo do metabolismo em diferentes condições experimentais usando uma técnica centrada em caminhos metabólicos (*pathway-centered approach*). Como explicado a seguir, esse método é uma extensão de uma técnica previamente publicada, em que métricas de caminhos de sinalização celular são utilizadas para classificação de células humanas em cancerosas e não-cancerosas.

A ideia do estudo do metabolismo *in silico* já foi apresentada no capítulo anterior onde tratamos da técnica de FBA e da nova proposta de uma função objetivo. Discutiremos brevemente duas extensões do método FBA que são interessantes e contribuem para facilitar o entendimento da técnica deste capítulo baseada em caminhos metabólicos.

O método RFBA (*Regulatory Flux Balance Analysis*) (COVERT; SCHILLING; PALSSON, 2001) é baseado em uma versão aprimorada de análise de balanço de fluxos e foi um dos primeiros métodos que busca-

ram a integração de informações metabólicas e regulatórias dentro da mesma estrutura de modelagem. Em RFBA os fluxos são restringidos a zero no caso dos mecanismos regulatórios (tanto fatores de transcrição quanto presença de metabólitos específicos no citoplasma) que controlam este fluxo não estarem ativos. O método permite simular situações metabólicas que dependam de variáveis regulatórias. É um método importante ainda hoje, e apresenta grande potencial de estudos, visto que permite uma modelagem *exata* das interações regulatórias. Todavia, a sua maior vantagem é também a sua maior deficiência, pois é um método pouco prático, dado o altíssimo número de interações regulatórias que devem ser pesquisadas e catalogadas antes de serem incorporadas manualmente ao modelo. Mesmo para subredes metabólicas menores com funcionalidade específica, como no caso apresentado em (RAMAN; RAJAGOPALAN; CHANDRA, 2005), a complexidade de modelagem pode ser impraticável. Adicionalmente, o método também não utiliza de forma explícita nenhum conhecimento de alto nível sobre a estrutura de caminhos metabólicos para as predições.

Em (CHANDRASEKARAN; PRICE, 2010), é apresentada uma técnica denominada PROM (*Probabilistic Integrative Modeling of Genome-Scale Metabolic and Regulatory Networks*), a qual foi apresentado como uma alternativa ao RFBA. A técnica utiliza FBA para caracterização metabólica dos dados, mas o faz apenas no contexto do estudo de taxas de crescimento. De acordo com os autores, esse é o primeiro esforço de integrar redes metabólicas e regulatórias em uma mesma técnica de maneira probabilística. No modelo PROM, a análise de distribuição de fluxos é utilizada para modelar matematicamente o metabolismo apenas com a matriz estequiométrica das reações metabólicas do organismo, ao passo que restrições regulatórias são obtidas a partir de dados de alta densidade de forma automatizada. O método PROM utiliza dados de experimentos de micromatrizes para estimar probabilidades de ativação de um gene a partir dos dados de expressão de mRNA. Embora o método consiga capturar o comportamento geral da célula em escala genômica, não utiliza nenhuma informação ou conhecimento sobre estruturas de caminhos metabólicos e, portanto, produz resultados apenas na forma de variações em genes isolados. Embora as predições sejam de natureza probabilística, e portanto mais robustas à presença de ruído e erros de modelagem, a estimação das probabilidades para cada gene não é baseada em nenhuma distribuição de probabilidade conhecida.

Embora nenhuma informação sobre caminhos metabólicos tenha sido utilizada em PROM ou RFBA, certos autores já estudaram ca-

minhos metabólicos individuais com FBA em um esforço de analisar o metabolismo em um nível mais alto. Em geral, as análises são realizadas com FBA incluindo dados experimentais, de forma similar ao método desenvolvido no capítulo anterior. As técnicas estudam um único caminho específico por vez, utilizando uma versão reduzida da rede metabólica (VARMA; PALSSON, 1994; RAMAN; RAJAGOPALAN; CHANDRA, 2005). Ainda outros métodos obtiveram, com sucesso, razoáveis predições metabólicas com FBA e uma rede metabólica em escala genômica. Porém isso foi feito mais uma vez sem incorporar nenhuma informação sobre a estrutura dos caminhos metabólicos subjacente (BESTE et al., 2007; COLIJN et al., 2009; SHLOMI et al., 2008; YIZHAK et al., 2010).

Na continuidade apresentaremos uma adaptação do método PathOlogist (EFRONI; SCHAEFER; BUETOW, 2007) para estudo do metabolismo. O PathOlogist pode ser visto como um bom compromisso entre os dois métodos, PROM e RFBA, discutidos acima, pois utiliza dados experimentais para estimar um modelo paramétrico para a expressão genética (transcriptômica), inclui informação de regulação genética e ainda permite o uso de conhecimento *a priori* sobre caminhos metabólicos provenientes de bases de dados. O método original foi apresentado no contexto de caminhos de sinalização celular apenas, com o objetivo de classificar fénotipos de câncer. Aqui, o método é estendido ao incorporar um modelo metabólico e informação sobre as enzimas responsáveis por catalisar cada uma as reações bioquímicas metabólicas individuais. Alguns outros métodos já foram publicados e estão disponíveis para o estudo de fenótipos com base em métricas de caminhos, tais como o *Signaling Pathway Impact Analysis* (SPIA) e o *Paradigm* (VARADAN et al., 2012), mas estes são mais específicos e direcionados ou para caminhos de sinalização celular ou para pesquisas do câncer humano, e a sua extensão para estudo do metabolismo em outros contextos ou com outros organismos não é simples.

O método PathOlogist, como originalmente desenvolvido, é específico para estudo de fenótipos de câncer em humanos, porém neste trabalho apresentamos a interessante extensão para estudo de atividade metabólica de organismos, mais especificamente do *Mycobacterium tuberculosis*. Entretanto, o método é flexível o suficiente e pode ser usado com qualquer outro organismo para o qual uma rede metabólica possa ser construída, mesmo incompleta, o que atualmente é uma atividade relativamente trivial, ainda que laboriosa. Com o método é possível estudar apenas um único caminho metabólico específico de interesse ou um subconjunto de caminhos a critério do investigador. Porém não

há impedimento, a não ser de complexidade computacional, para que mesmo todos os caminhos que compõem o metabolismo de um organismo sejam estudados. O nosso *objetivo* com este trabalho foi o de desenvolver e apresentar um método simples para análise de dados de transcriptômica com foco na estrutura dos caminhos metabólicos e cujos resultados fossem de fácil interpretação pelo usuário.

A *justificativa* para a escolha do método PathOlogist foi baseada na sua simplicidade e flexibilidade, permitindo ampliações, variações e alterações de forma relativamente simples, além de produzir métricas que avaliam diretamente caminhos metabólicos a partir de medidas de micromatrizes de DNA e/ou espectrometria de massa. A extensão do método, que apresentaremos a seguir, é uma boa alternativa ou complemento ao método de análise de balanço de fluxos (FBA).

Um dos usos para o método que será apresentado é como uma ferramenta auxiliar no desenvolvimento de compostos antibióticos para combater organismos patogênicos, como o *Mycobacterium tuberculosis*. Finalmente, é importante compreender que o método proposto não é um substituto para o método PathOlogist original, visto que os dois possuem objetivos diferenciados. Um método para o estudo do metabolismo é útil para o pesquisador que deseja entender quais mecanismos metabólicos são alterados em função da alterações ambientais externas e em função da reprogramação celular (via caminhos de sinalização) daí decorrente. No caso do Mtb, devido à sua alta capacidade de desenvolver resistência aos antibióticos atuais, é desejável conhecer as alterações metabólicas que são subjacentes ao desenvolvimento da resistência, pois esta pode ser a chave para o desenvolvimento de novas drogas antibióticas mais eficazes.

4.2 VISÃO GERAL DO MÉTODO PATHOLOGIST

O método PathOlogist original foi detalhadamente descrito e revisado previamente em outras publicações (EFRONI; SCHAEFER; BUETOW, 2007; EFRONI et al., 2011; VARADAN et al., 2012). O PathOlogist surgiu da necessidade de ferramentas que pudessem realizar uma análise quantitativa de caminhos de sinalização celular já identificados. Embora diversas ferramentas e técnicas estejam disponíveis para estimar, criar, inferir e visualizar estruturas de caminhos (e.g. Cytoscape (SHANNON et al., 2003), GenePath (ZUPAN et al., 2003), GeneNet (OPGENRHEIN; STRIMMER, 2007)), existem poucas técnicas disponíveis para redes metabólicas e que realmente utilizem essa informação combinada

com dados experimentais de alta densidade. O método PathOlogist usa dados de expressão de mRNA obtidos de experimentos de micromatrizes de DNA e calcula duas métricas de caminhos, a *atividade* e a *consistência* de um caminho de sinalização celular. Isso é feito para cada caminho que o investigador deseja estudar. Essas métricas de alto nível substituem os dados originais de expressão em todas as análises subsequentes, como agrupamento hierárquico de caminhos, visualização de componentes e estrutura de caminhos, correlações das atividades de diferentes caminhos, classificação de amostras clínicas, e análise estatística da associação entre padrões de atividade de caminhos e fenótipos ou padrões clínicos observados. O potencial do método reside no fato de ele transformar grandes conjuntos de dados de expressão genética de alta dimensionalidade em descritores quantitativos com menor dimensão e de mais fácil interpretação do comportamento no que diz respeito aos caminhos regulatórios (metabólicos) como mostra a Figura 21. Nessa figura mostramos uma diagrama em blocos mostrando o objetivo do método PathOlogist, que é o de transformar conjuntos de dados de expressão de mRNA (obtidos por exemplo com experimentos de micromatrizes de DNA de alta densidade) em escores probabilísticos para caminhos de sinalização celular. Duas vantagens do método são a redução da dimensionalidade dos dados (pois, em geral, o número de genes i no genoma é muito maior que o número de caminhos k de sinalização) e a maior facilidade de interpretação dos resultados, pois informação sobre a atividade de um caminho de sinalização representa muito mais informação para o investigador do que o valor de expressão de um conjunto de genes individuais. Observa-se na figura que a probabilidade do estado de ativação de cada gene é uma informação necessária no mapeamento realizado.

Em sua forma original, o método PathOlogist foi projetado especificamente para estudar caminhos de sinalização de células cancerosas (VARADAN et al., 2012), em que cada caminho é definido como um conjunto conectado de interações gene-gene, e em que uma interação é composta de um conjunto de genes de entrada (*input genes*) que podem ativar ou inibir um conjunto correspondente de genes de saída (*output genes*), como no exemplo mostrado na Fig. 22.

No método PathOlogist, escores de atividade e consistência são calculados para cada interação. No exemplo da Fig. 22, o escore de atividade da interação é calculado da seguinte forma:

$$\text{Atividade} = p(A) \cdot p(B) \tag{4.1}$$

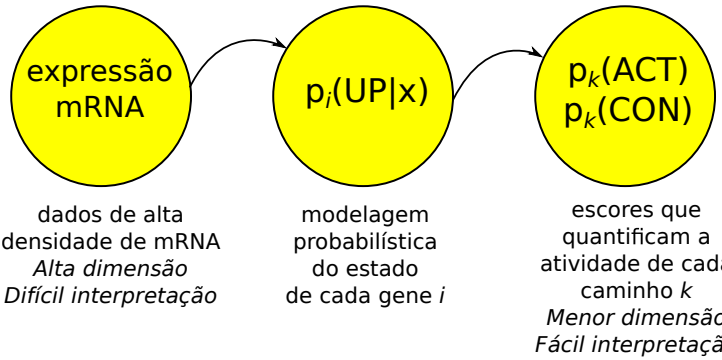

Figura 21 – O método PathOlogist calcula escores de atividade e consistência para um caminho de sinalização celular a partir de dados de expressão genética, transformando grandes conjuntos de dados de mRNA em métricas numéricas representativas do funcionamento celular em termos das fluxo de 'mensagens' moleculares que ocorrem dentro do ambiente celular. Os escores de atividade e consistência são calculados usando a probabilidade do estado do gene (ativo ou inativo) dado o valor de expressão genética x observado para cada um dos i genes participantes de cada caminho k. *Fonte: Adaptado de (EFRONI et al., 2011).*

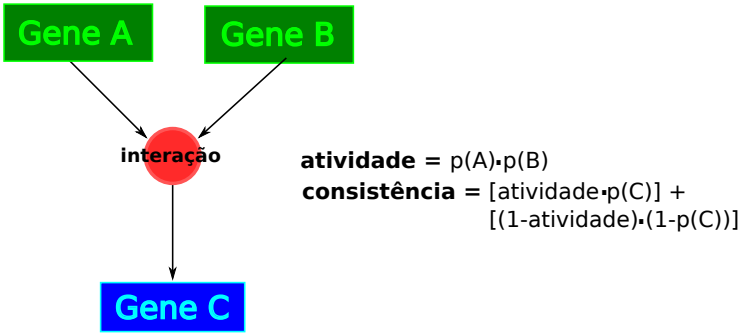

Figura 22 – Definição de uma interação no método PathOlogist original. A ação combinada de um conjunto de genes de entrada (A e B) pode inibir ou ativar um número de genes de saída (gene C neste exemplo). *Fonte: Adaptado de (EFRONI et al., 2011).*

em que $p(A)$ é a probabilidade de que o gene de entrada A esteja em um estado ativado, que chamaremos de *UP state*, e representa um estado de alta atividade ou de alto nível de expressão dentro da célula, sendo $p(B)$ o equivalente para o gene B. O escore de consistência por sua vez é calculado como (EFRONI; SCHAEFER; BUETOW, 2007):

$$\text{Consistência} = [\text{Atividade} \cdot p(C)] + [(1 - \text{Atividade}) \cdot (1 - p(C))] \quad (4.2)$$

em que $p(C)$ é a probabilidade de que o gene C esteja também em um estado de alta expressão e atividade dentro da célula. O escore de consistência é uma medida de quão bem os dados observados concordam com a estrutura de rede subjacente do caminho. Finalmente, para gerar o escore para todo um caminho, os escores de todas as interações neste caminho são combinadas em um valor médio, de forma que alterações em qualquer parte do caminho tenham o mesmo efeito geral. Isso, entretanto, não significa que não se possa analisar interações específicas dentro do caminho, algo que pode facilmente ser feito apenas observando os escores individuais de cada interação.

Como a quantidade de dados obtidos de medições como micromatrizes de DNA, espectrometria de massa e outras técnicas mais avançadas já discutidas na introdução, é cada vez maior (aumentando de forma drástica e paulatinamente a razão p/n entre dimensão dos dados e número de amostras), qualquer trabalho de modelagem precisa em primeiro lugar buscar descritores de dimensionalidade reduzida que sejam resumos úteis dessa grande quantidade de dados. Tais descritores permitirão uma melhor caracterização das amostras clínicas do que é possível ter usando-se unicamente dados crus provenientes dos experimentos.

Em segundo lugar, nem sempre os dados desejados são os dados que podem ser obtidos. Por exemplo, é fácil obter dados de expressão genética, mas não tão simples obter dados de fluxo metabólico. Entretanto, sabe-se que estas duas maneiras diferentes de visualizar o comportamento celular estão intimamente ligadas e possuem informações correlacionadas e possivelmente sobrepostas. Qualquer modelo portanto necessita também ser capaz de transformar a representação dos dados, retirando informação de um tipo de dado a partir de outros tipos de dados diferentes (e.g. extrair informação sobre fluxo metabólico a partir de dados de proteômica como mostrado no capítulo anterior).

Muita pesquisa foi realizada na área da biologia computacional com o intuito de encontrar descritores (e.g. biomarcadores) como subconjuntos de genes ou de proteínas com variação na expressão acima

ou abaixo de determinado limiar. Todavia, os resultados obtidos dessa forma são em geral desapontadores, não produzindo o tão desejado resultado de indicar de forma inequívoca alvos racionais para novos compostos antibióticos. Isso ocorre pois estes agrupamentos de genes ou proteínas individuais não são suficientes para esclarecer os mecanismos celulares, sejam eles metabólicos, regulatórios ou de sinalização. Em muitos casos, diferentes estudos podem produzir resultados diferentes e mesmo contraditórios para situações experimentais similares (MACHADO; HERRGARD, 2014). A dificuldade está em que apenas uma lista de genes que se expressam diferencialmente em diferentes condições experimentais *não* revela automaticamente os mecanismos moleculares, regulatórios ou metabólicos subjacentes responsáveis pelo fenótipo observado. São necessários modelos mais sofisticados, que produzam descritores de mais alto nível e de mais fácil interpretação, descritores que estejam mais próximos do fenótipo que os dados crus, como o método desenvolvido no presente capítulo.

Na próxima seção é detalhada a metodologia proposta, o Pathmod, para estender o método PathOlogist para o estudo das redes metabólicas e da atividade de caminhos metabólicos.

4.3 PATHOLOGIST MODIFICADO PARA METABOLISMO

Assim como em redes de sinalização celular, o metabolismo também pode ser descrito em termos de uma rede (NEWMAN, 2010), como discutido no capítulo anterior. Nessa rede, os nós são metabólitos conectados por reações metabólicas, geralmente catalisadas por um conjunto de enzimas. Para estudarmos metabolismo, propomos inicialmente usar não um conjunto de genes de sinalização como no método PathOlogist original, mas um conjunto de enzimas com proeminente papel metabólico, e que em última instância catalisam as reações metabólicas e determinam o fluxo dessas reações.

Ao introduzirmos essa modificação podemos calcular tanto escores de atividade quanto de consistência para um grupo de genes que codificam proteínas com atividade enzimática. Esses escores individuais para cada enzima, quando combinados de forma apropriada, permitem obter *escores de reações metabólicas*, utilizando expressões que descrevem a atividade catalítica combinada de um conjunto de enzimas sobre uma reação. Os escores de reações são finalmente combinados em um valor médio, que representa o escore de um caminho metabólico. Para gerar estes escores de caminhos, os escores de todas as reações

metabólicas que fazem parte de um caminho (essa informação pode ser obtida de bases de dados metabólicas específicas para o Mtb como TBDB, Tuberculist ou KEGG) são usados. Diferentemente do método PathOlogist original, em que o conceito de caminho representa um conjunto de interações entre produtos genéticos, no método Pathmod proposto um *caminho* é a noção amplamente aceita de uma sequência de reações metabólicas catalisadas por enzimas e atuando de forma coordenada para atingir um determinado objetivo metabólico na célula, seja catabólico, anabólico ou de transporte.

Na alteração proposta para estudo do metabolismo, os escores de atividade e consistência de um determinado caminho metabólico são calculados a partir dos escores das reações bioquímicas que compõem o caminho. Os escores das reações por sua vez são calculados a partir dos escores das enzimas que catalisam uma determinada reação. Esses escores de enzimas são calculados a partir das probabilidades dos fatores de transcrição que regulam a atividade transcripcional das enzimas. Inicialmente as probabilidades de genes estarem em um estado ativo (UP) são calculadas a partir de um modelo probabilístico de atividade, seja uma única função densidade de probabilidade (PDF) gama ou uma mistura de gamas como descrito a seguir. Os parâmetros desse modelo são estimados a partir de um conjunto de treinamento independente. É possível utilizar dados de proteômica caso esses experimentos estejam disponíveis, porém no presente trabalho apenas dados transcripcionais foram utilizados. A Figura 23 apresenta um fluxograma geral do método Pathmod.

O método faz duas importantes suposições. Primeiramente, assume-se que a expressão genética para um gene com estado basal único segue uma distribuição gama (EFRONI; SCHAEFER; BUETOW, 2007). Assume-se que a expressão genética de um gene com caráter bimodal é dada por uma mistura de duas distribuições gama, uma para gene inativo DN e outra para gene ativo UP (EFRONI; SCHAEFER; BUETOW, 2007). Em segundo lugar assume-se que a atividade de fatores de transcrição é independente da atividade de enzimas e, portanto, probabilidades de ativação de uma enzima podem ser obtidas pela multiplicação das probabilidades dos seus genes reguladores, e da mesma forma a probabilidade de uma reação metabólica em relação às suas enzimas catalisadoras.

Quanto à escolha de uma distribuição gama para modelagem dos dados de transcriptômica, é importante notar que a modelagem de dados de expressão gênica com distribuições paramétricas é uma questão em aberto (ver (THOMAS et al., 2010) e (LEE, 2004), seção 13.1). No

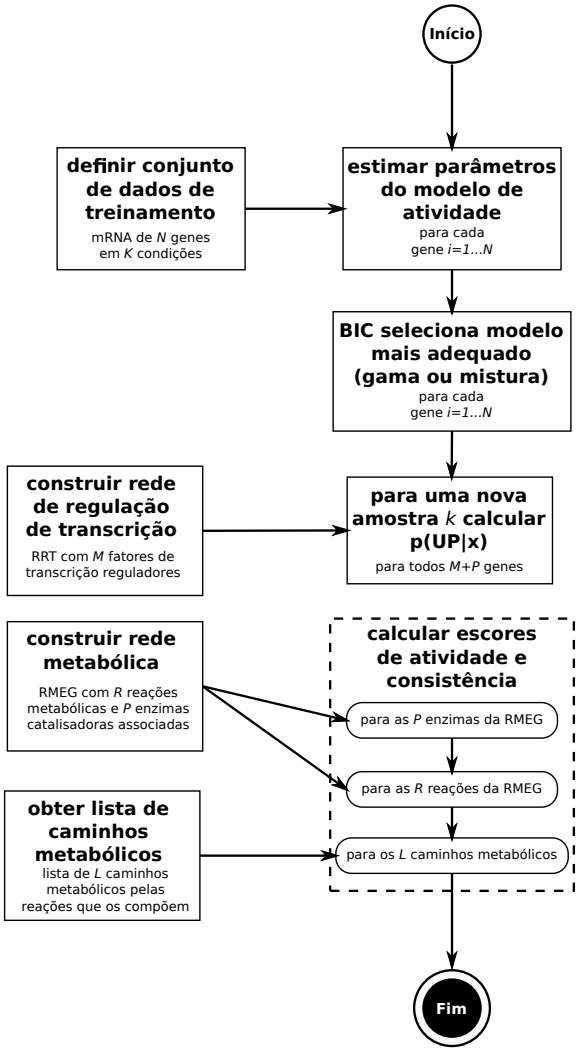

Figura 23 – Fluxograma dos passos de cálculo do método Pathmod. Na esquerda, os blocos indicam fontes de dados necessárias: dados de treinamento; rede de regulação transcripcional; rede metabólica com reações metabólicas e enzimas; listagem de caminhos metabólicos e reações que compõem o caminho. Os blocos à direita indicam os passos de cálculo: estimação do modelo de atividade para cada gene; seleção do modelo para cada um dos N genes; cálculo das probabilidades de estado ativado para os M fatores de transcrição e P enzimas envolvidos nos L caminhos metabólicos; cálculo em seqüência dos escores Pathmod para as P enzimas, R reações e L caminhos metabólicos. *Fonte: Elaborada pelo autor.*

capítulo 2 assumimos que o vetor de expressão transcriptômica *de todos os genes do organismo em uma única condição experimental* (dimensão p do vetor aproximadamente 4000 para o *Mycobacterium tuberculosis*) poderia ser modelado por uma distribuição normal multivariada(BALDI; LONG, 2001), pois essa distribuição apresenta vantagens quando o objetivo do modelo é a geração de dados sintéticos.

Neste capítulo assume-se, da mesma forma como em (EFRONI; SCHAEFER; BUETOW, 2007), uma distribuição gama *univariada* para os dados de transcriptômica, porém não mais para a expressão de todos os genes em uma única condição, mas para a expressão *cada gene em todas as condições experimentais disponíveis*. Essa é a distribuição geralmente considerada quando da modelagem da distribuição marginal da expressão de um gene observado em diversas condições experimentais (LEE, 2004).

O método necessita de 4 fontes de informação para cálculo dos escores. Em geral, essas informações estão disponíveis em bases de dados públicas e podem ser acessadas de forma automatizada. Para estimação dos parâmetros dos modelos de atividade dos genes é necessário o uso de um conjunto de dados de expressão genética (transcriptômica) que seja independente dos dados a serem estudados a partir desses modelos. No presente trabalho utilizou-se o conjunto de dados Boshoff publicado e descrito em (BOSHOFF et al., 2004) e apresentado brevemente mais adiante. Os parâmetros são estimados por máxima verossimilhança e, no caso de uma mistura, utilizando-se o algoritmo EM (*expectation-maximization*) (BISHOP, 2007). São necessárias também uma rede metabólica em escala genômica (RMEG) contendo um número de reações metabólicas de interesse que possam ser agrupadas em caminhos metabólicos coerentes e as enzimas que catalisam cada uma das reações. As reações desse modelo devem ser agrupadas em caminhos em metabólicos, as quais serão usadas no último passo do método para encontrar os escores de atividade e determinar a consistência para cada um dos caminhos metabólicos. Há ainda a necessidade de informação sobre a rede de regulação do organismo, contendo todos os genes reguladores (i.e. fatores de transcrição) importantes para o organismo e as enzimas reguladas por estes genes.

Embora a quantidade de informação e dados necessária para o método seja bastante significativa (um problema associado com todo modelo em escala genômica que utiliza dados de alta densidade e deseja incorporar fontes de dados diversas), a maior dificuldade está na criação de métodos de busca automática desses dados, visto que uma grande quantidade de informação (embora muitas vezes ainda não vali-

dada ou constantemente atualizada) está disponível em artigos e bases de dados com acesso público. Por exemplo, para este trabalho a rede transcripcional foi obtida dos artigos (BALÁZSI et al., 2008; SANZ et al., 2011) e complementada com dados mais atualizados disponíveis na base de dados (TB Database, 2015b) para o *Mycobacterium tuberculosis*. A rede metabólica utilizada está disponibilizada pelos autores de (BESTE et al., 2007) e foi atualizada com algumas reações importantes de processos de síntese de parede celular descritos em (COLIJN et al., 2009). A listagem de caminhos metabólicos e das reações participantes e algumas enzimas catalisadoras não disponíveis em (BESTE et al., 2007) foi obtida da base de dados KEGG (Kanehisa Labs, 2015). Na próxima seção descreve-se em detalhes o procedimento de cálculo dos escores dos caminhos metabólicos.

4.4 ESCORES PARA CAMINHOS METABÓLICOS

Escores de caminhos são calculados a partir de dados de expressão genética para uma condição experimental desejada (e.g. um canal de uma micromatriz de DNA). Para isso, um primeiro passo necessário é o cálculo da probabilidade de cada gene estar em um estado ativado dados o valor de expressão observado. Chamaremos o estado ativado de condição *UP*, e o valor de expressão genética de x:

$$P_{ij}(\text{State} = UP|x_{ij}) \tag{4.3}$$

em que x_{ij} é o valor de expressão mRNA observado para o gene i na condição experimental j. Quando dizemos "a probabilidade de um gene estar em um estado ativado (estado *UP*)", queremos dizer a probabilidade que temos de encontrar o gene em um estado de alta atividade dentro da célula em uma determinada condição experimental. No caso de um fator de transcrição (FT), isso significa a probabilidade de que os alvos desse FT tenham sua expressão ativada, e no caso de uma enzima, a probabilidade de que a reação metabólica catalisada apresente um maior fluxo metabólico. Geralmente uma condição experimental corresponde a uma única micromatriz de DNA, e se experimentos replicados estão disponíveis calcula-se o valor médio das diversas repetições (EFRONI; SCHAEFER; BUETOW, 2007). Para simplificar a notação, nos cálculos subsequentes eliminamos o índice i de identificação do gene, e subentende-se que o modelo probabilístico é estimado para cada gene separadamente. Visto que o índice j também não é relevante para o desenvolvimento do modelo de atividade do gene, este também é supri-

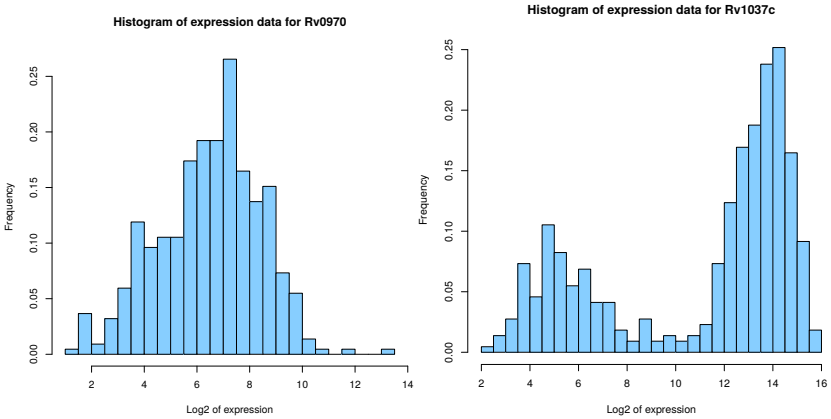

Figura 24 – Histogramas de dados de expressão genética para todas as amostras de um conjunto de dados (BOSHOFF et al., 2004), para dois genes selecionados. A imagem à esquerda é o $\log_2$ da expressão do gene *Rv0970c*, apresentando um comportamento aproximadamente unimodal (gene está sempre expresso no citoplasma com determinado nível basal constante). O histograma à direita mostra comportamento bimodal (bi-estado) para o gene *Rv1037c*, com um distinto limiar entre o estado de baixa expressão (DN) e o estado de alta expressão (UP). *Fonte: Elaborada pelo autor.*

mido da notação. A seguir apresentamos o método de estimação das probabilidades *UP* a partir de dados de expressão como em (EFRONI; SCHAEFER; BUETOW, 2007), que é a mesma técnica utilizada no modelo Pathmod proposto. A expressão para um único gene através de um conjunto de micromatrizes de DNA pode, em geral, ser considerada como unimodal ou bimodal dependendo do seu histograma (KAERN et al., 2005). Essas duas possibilidades estão mostradas na Figura 24 para os genes *Rv0970c* e *Rv1037c*. Como não é possível saber de antemão qual distribuição é mais adequada para cada um dos 4000 genes para os quais um modelo probabilístico precisa ser estimado, estimam-se inicialmente dois modelos para cada gene: uma distribuição gama única e uma mistura de duas componentes gama, em que cada componente modela um dos possíveis estados do gene: *UP* ou *DN*. Embora seja possível modelar diretamente a expressão genética diferencial como é feito por exemplo em (LEWIN et al., 2006), optou-se por modelar a distribuição marginal dos genes para cada micromatriz de DNA separadamente, controle e tratamento, seguindo de forma mais próxima o modelo PathOlogist original (EFRONI; SCHAEFER; BUETOW, 2007). Outra

motivação para modelarmos as condições de controle e tratamento separadamente é porque essa modelagem permite demonstrar que o método pode distinguir não apenas tratamentos diversos, mas mesmo condições de controle das condições de tratamento correspondentes.

Embora diversas distribuições de probabilidade possam ser utilizadas para modelar a distribuição marginal da expressão genética (gama, log-normal, normal, etc.) (WIJAYA; HARADA; HORTON, 2008; NEWTON et al., 2004), a mais utilizada é a distribuição gama (KHALILI et al., 2007; EFRONI; SCHAEFER; BUETOW, 2007), pois é uma distribuição assimétrica com formato flexível e, portanto, adequada para modelagem da distribuição de variáveis aleatórias não-negativas assimétricas. Na estimação do nosso modelo Pathmod, essa foi a distribuição escolhida, por ser também é a distribuição utilizada no método original PathOlogist (EFRONI; SCHAEFER; BUETOW, 2007). O tópico sobre qual modelo probabilístico deve ser utilizado na modelagem da expressão genética é importante mas também relativamente complexo (SEGAL et al., 2001; PAULSSON, 2005). Dessa forma, visto que modelos gama e de misturas de distribuições gama já têm sido utilizados com sucesso na modelagem de diversos sistemas biológicos (MAYROSE; FRIEDMAN; PUPKO, 2005; KELES, 2007) e são uma escolha razoável para distribuições assimétricas, optou-se por deixar semelhante discussão para um possível futuro trabalho, visto o foco deste trabalho ser o desenvolvimento de uma metodologia para estudo de metabolismo com métricas de caminhos metabólicos e não a seleção da PDF mais adequada para modelagem de expressão genética.

A estimação dos parâmetros para ambos os tipos de modelo (unimodal e bimodal) é realizada respectivamente por máxima verossimilhança para o caso unimodal e com o algoritmo EM (*expectation-maximization*), no caso da mistura. A estimação da distribuição unimodal pode ser obtida facilmente em R com a função `fitdistr()` do pacote MASS. No caso da estimação da mistura, visto que os pacotes oferecem apenas a estimação de misturas de distribuições gaussianas, foram desenvolvidas as expressões apresentadas no Apêndice A, programando-se o algoritmo EM iterativo em um script R.

Os parâmetros para ambos os modelos são estimados utilizando o conjunto de dados Boshoff (BOSHOFF et al., 2004). Esse é um extenso conjunto de dados de micromatrizes de DNA com o *Mycobacterium tuberculosis*. O conjunto de dados Boshoff compreende 437 micromatrizes de DNA de dois canais com medidas de expressão para cada um dos aproximadamente 4000 genes do *Mycobacterium tuberculosis* para 75 tratamentos em diversas condições experimentais. Cada micromatriz é

composta de um canal com a condição de controle (CTL/Cy3), e um outro canal com a condição de tratamento (TRT/Cy5), de forma que $437*2 = 874$ observações estão disponíveis. O conjunto inclui tratamentos com diversos compostos antibióticos conhecidos como a isoniazida, etambutol e rifampicina, e também culturas em outras condições de estresse como depleção de oxigênio, meios acídicos (baixo pH) e limitação de nutrientes. Para uma listagem completa dos diversos tratamentos disponíveis ver Apêndice B. Todos os dados dos experimentos estão disponíveis publicamente na forma $\log_2$ dos valores de expressão. Notamos que no modelo original do PathOlogist, os parâmetros do modelo probabilístico de atividade genética são estimados a partir dos mesmos dados usados nos cálculos ulteriores. Visto que este procedimento está sujeito a críticas fundadas sobre ser um procedimento estatístico razoável, já que as estimativas obtidas do modelo poderão ser polarizadas em direção aos dados usados no passo de treinamento, na extensão Pathmod que propomos aqui sugerimos estimar os parâmetros utilizando sempre um conjunto de dados independente, o qual não inclui nenhuma micromatriz que será usada posteriormente para o cálculo dos escores dos caminhos. Como é desejável ter escores para todos os tratamentos incluídos no conjunto de dados Boshoff, essa estimação foi então realizada uma vez para cada tratamento, semelhante ao método *leave-one-out*. Utilizando o método Pathmod, foi possível transformar cada uma das micromatrizes contendo 4000 valores de expressão genética de difícil interpretação em escores de atividade metabólica com dimensão 82, oferecendo uma interpretação alternativa e possivelmente mais rica dos dados. Essas métricas são calculadas para cada caminho e são como "resumos" de alto nível (i.e. no nível de caminhos metabólicos) do estado biológico da célula em uma determinada condição experimental. O modelo de dados para a expressão genética no caso de uma única distribuição gama (i.e. histograma unimodal) é dada por:

$$p(x) = f(x|a,b) = \frac{x^{a-1}}{\Gamma(a)b^a}e^{-x/b} \tag{4.4}$$

em que estamos implicitamente assumindo que esse é o comportamento estatístico da expressão do gene quando o gene está em seu estado ativado (UP). Essa é exatamente a mesma suposição feita em (EFRONI; SCHAEFER; BUETOW, 2007). Em nosso modelo, todavia, fazemos uso dessa suposição de maneira menos estrita, como apresentaremos em breve. O segundo modelo, o modelo de mistura de gamas, é dado por:

$$p(x|\mathrm{S}) = \pi_u \mathcal{G}(a_u, b_u) + \pi_d \mathcal{G}(a_d, b_d) \tag{4.5}$$

em que $\mathcal{G}(a_k, b_k) = \frac{x^{a_k-1}}{\Gamma(a_k) b_k^{a_k}} e^{-x/b_k}$ é uma componente gama da mistura, e os coeficientes da mistura são indicados pelas variáveis π_u e π_d, com $\pi_u + \pi_d = 1$ e $p(x|\mathrm{S})$ é a probabilidade da observação do valor de expressão x dado que o estado do gene seja S. Essas variáveis são estimadas juntamente com os parâmetros da mistura com o método EM e representam respectivamente as probabilidades *a priori* de que o gene é observado em um dos dois estados possíveis:

$$\begin{aligned} \pi_u &= P(\mathrm{S} = UP) \\ \pi_d &= P(\mathrm{S} = DN) \end{aligned} \tag{4.6}$$

Notamos também que cada termo gama na Eq. (4.5) representa a verossimilhança dos dados dado o estado, de forma que é possível reescrever:

$$p(x) = P(\mathrm{S} = UP)p(x|\mathrm{S} = UP) + P(\mathrm{S} = DN)p(x|\mathrm{S} = DN) \tag{4.7}$$

Observa-se que a probabilidade desejada $P(\mathrm{State} = UP|x)$ no caso do modelo de mistura de gamas é dada por:

$$p(\mathrm{S} = UP|x) = \frac{P(\mathrm{S} = UP)p(x|\mathrm{S} = UP)}{p(x)} \tag{4.8}$$

em que $p(x) = \pi_d \mathcal{G}(a_d, b_d) + \pi_u \mathcal{G}(a_u, b_u)$. No caso de uma distribuição gama unimodal, no método PathOlogist original (EFRONI; SCHAEFER; BUETOW, 2007), a probabilidade *a posteriori* desejada $P(\mathrm{State} = UP|x)$ é simplesmente ajustada para o valor 1.0 em todos os cálculos subsequentes[1], descartando completamente a PDF gama estimada e assumindo implicitamente que na situação em que o gene apresenta comportamento unimodal aproximado (ver Fig. 24 esq.), o seu nível basal de expressão sempre corresponde ao estado totalmente ativado.

Embora não seja biologicamente razoável que o gene nunca entre em seu estado ativado (UP), assumir que o seu estado é sempre totalmente ativado também apresenta dificuldades. Essa suposição desconsidera duas situações possíveis: (1) o conjunto de dados utilizado na estimação dos parâmetros das PDFs não continha um número significativo de condições experimentais em que o gene apresentaria expressão mais alta característica de um estado de super expressão (UP) distinto do estado de baixa expressão (o experimento não foi suficientemente compreensivo), e (2) a diferença dos valores de expressão entre os estados distintos UP e $DOWN$ para um determinado gene não apresenta magnitude sufiente para que o histograma apresente aparência bimodal.

[1] *Fonte*: código-fonte disponibilizado pelo autor do método PathOlogist.

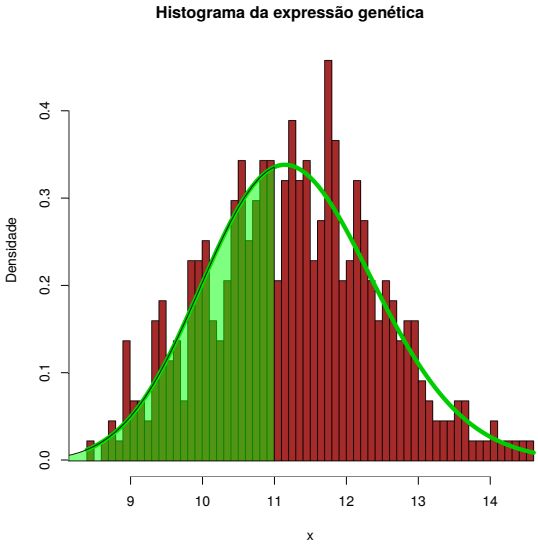

Figura 25 – Diferentemente do método PathOlogist original, no caso de densidades unimodais para modelagem dos dados de expressão, utilizamos a função distribuição cumulativa de probabilidade para o estado do gene, dado um valor de expressão x. A área verde sob a PDF ajustada determina a probabilidade do gene estar em um estado ativado (*UP*) dado o valor de expressão observado na micromatriz (na figura, $x = 11$). *Fonte: Elaborada pelo autor.*

De forma a evitar esse tipo de comportamento extremo, nós propomos estimar a probabilidade do estado *UP* state para o gene com base na sua cdf (*cumulative distribution function*), i.e. a integral da densidade gama estimada desde zero até o valor de expressão para o qual deseja-se determinar a probabilidade do estado, conforme apresentado na Figura 25.

As duas situações, densidade gama única e mistura de densidades gama, estão representadas na Figura 24 para dois genes específicos tomados do conjunto de dados Boshoff. Visto que alguns genes podem apresentar comportamento unimodal, enquanto outros se ajustarão melhor a uma densidade de probabilidade bimodal, a decisão de qual é o modelo mais apropriado para cada gene (uma densidade gama ou uma mistura de gamas) é feita com base no critério BIC (*Bayesian Information Criterion*). A métrica BIC (CONGDON, 2006) é um método simples para identificar automaticamente o melhor modelo dentre um conjunto

de possíveis modelos que deseja-se ajustar a um conjunto de dados, e é recomendado ao invés do AIC, pois apresenta uma menor tendência a superestimar o número de componentes da mistura (WIJAYA; HARADA; HORTON, 2008). Diversas referências estão disponíveis sobre a modelagem de dados de expressão genética e proteica e as complexidades associadas, que tratam, ainda que indiretamente, da identificação do número de componentes da mistura (KAERN et al., 2005; PAULSSON, 2005). Nesses trabalhos o objetivo é definir se um determinado gene é melhor caracterizado por níveis de expressão de baixa atividade ou alta atividade, ou apenas um nível de atividade. A complexidade desta discussão não permite que o tema seja abordado de forma mais detalhada neste trabalho. Dessa forma, e seguindo ainda as diretrizes tomadas no método PathOlogist original, a decisão sobre o número de componentes é realizada apenas com base no critério BIC. O critério é dado por:

$$BIC = -2 * (-\text{LogLik}) + \log(N) * K \tag{4.9}$$

em que LogLik é o valor da função log-verossimilhança para o conjunto de parâmetros, estimados pelo método recursivo EM, que maximiza o valor da função, N é o número de observações disponíveis para a estimação, que neste caso é o número total de amostras de micromatrizes em que a expressão do gene foi medida (437 2-channel micromatrizes no caso de ser utilizado todo o conjunto de dados Boshoff), e K é o número de parâmetros no modelo, o qual é igual a dois no caso da densidade única e igual a seis no caso da mistura.

Utilizando a expressão Eq.(4.9) calculou-se para cada gene o valor do critério para cada modelo possível (distribuição gama e mistura de duas gamas) e selecionou-se o modelo que apresentava o melhor ajuste aos dados escolhendo aquele com o menor valor de BIC. Como um exemplo, o gene *Rv1037c* estimado com todo o conjunto de dados, para o qual a mistura de gamas claramente é o melhor modelo (ver Figura 26), os valores de BIC obtidos para cada modelo são:

$$\begin{aligned} \text{BIC}_{uni} &= 2541.69 \\ \text{BIC}_{mix} &= 2125.97 \end{aligned} \tag{4.10}$$

Os dois modelos são apresentados sobrepostos aos histogramas na Figura 26. Nesta figura, o modelo da mistura de gamas é nitidamente o mais adequado. A componente da mistura à esquerda (vermelho, estado DN) é dada por $\mathcal{G}(a_d, b_d)$ com parâmetros $a_d = 36.03$ and $b_d = 6.6$. A PDF à direita (verde, estado UP) é $\mathcal{G}(a_u, b_u)$ com parâmetros $a_u = 130.7$ and $b_u = 9.7$. Os coeficientes da mistura estimados são $\pi_d = 0.31$

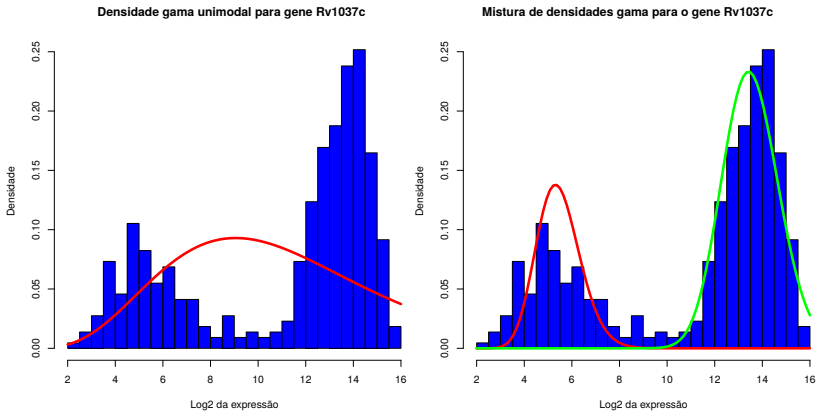

Figura 26 – Histogramas dos dados de expressão genética e funções densidade de probabilidade ajustadas usando o conjunto de dados Boshoff para o gene *Rv1037c*. Histograma na esq. é o modelo de uma única distribuição gama conforme Eq. (4.4). Histograma na dir. é o modelo de mistura de gamas conforme Eq. (4.5). *Fonte: Elaborada pelo autor.*

e $\pi_u = 0.69$. A PDF única (modelo menos adequado aos dados) foi estimada diretamente por máxima verossimilhança. A mistura de gamas é estimada com o método iterativo *expectation-maximization* (EM) algorithm (BISHOP, 2007) que maximiza a função log-verossimilhança. Para evitar dificuldades associadas à convergência do algoritmo EM, os parâmetros das componentes gama são inicializados utilizando-se o algoritmo de agrupamento k-means (BISHOP, 2007; THEODORIDIS; KOUTROUMBAS, 2008) para agrupar os dados em dois grupos, em seguida ajustando-se uma distribuição gama para cada agrupamento de dados separadamente. Observou-se a partir dos resultados com a métrica BIC, para o conjunto de dados Boshoff, dos 3924 genes do *Mycobacterium tuberculosis* disponíveis na micromatriz, a expressão de 2220 genes é mais adequadamente modelada por uma distribuição unimodal (uma PDF gama), ao passo que para os 1704 genes restantes, sua expressão genética é melhor descrita por uma mistura de distribuições gama.

Tendo estimado as PDFs para o estado de ativação dado o valor de expressão, é possível calcular para um novo conjunto de dados de micromatrizes de DNA, para cada gene i (retângulos azuis e verdes na Figura 27) e condição experimental j, a probabilidade $P_{ij}(\mathrm{S} = UP|x_{ij})$ de que um gene i específico está no seu estado ativado (UP) dados o valor de expressão observado na micromatriz j. Embora nesse método

seja utilizada apenas a probabilidade de um gene estar em seu estado ativado, caso seja de interesse utilizar a probabilidade de um gene estar em seu estado inibido (DN), visto a mistura ser de apenas duas componentes e o número de estados assumidos apenas dois, basta usar o valor $1 - P(UP|x)$.

Após termos obtido as probabilidades para cada fator de transcrição regulador, e assumindo independência das suas PDFs, estamos prontos para calcular os escores de atividade e consistência das enzimas utilizando a probabilidade de ativação dos fatores de transcrição, da mesma forma como mostra a Figura 22. Multiplicando as probabilidades dos genes de entrada calcula-se a probabilidade de materialização da interação como sendo a probabilidade conjunta de todos os genes de entrada, i.e. multiplicam-se as probabilidades UP de todos os fatores de transcrição reguladores para obter o escore de atividade das enzimas reguladas. A relação regulador/regulado entre fatores de transcrição e enzimas é dada pela informação contida na rede de regulação de transcrição (RRT), conforme explicado acima no início desta seção.

Os escores de consistência são calculados de forma similar ao método PathOlogist original, em que os valores de expressão de enzimas são usados para avaliar a consistência entre as probabilidades de atividade obtidas diretamente dos dados da micromatriz e o escore de atividade calculado a partir dos fatores de transcrição. O cálculo da consistência das reações metabólicas, visto não estarem disponíveis valores de probabilidade de fluxo, é realizado de forma similar ao cálculo dos escores de atividade, como será explicado a seguir.

Apresentamos agora o método Pathmod por meio de um exemplo simplificado. Na Figura 27, apresentamos uma representação gráfica do método que ajudará a entender como é feito o cálculo dos escores Pathmod para o caminho metabólico de biossíntese da metionina (METHBio) indexado na base de dados KEGG sob o código *mtu00270*. Essa figura serve para descrever também as modificações em relação ao método PathOlogist original necessárias para estudo do metabolismo. No método original, há apenas genes que interagem passando 'mensagens' bioquímicas de sinalização. Não há fatores de transcrição, enzimas ou reações metabólicas. Visto ser mais simples entender o método iniciando pelo caminho metabólico, explicaremos a figura da direita para a esquerda. Na figura, o retângulo amarelo representa o caminho metabólico de biossíntese do amino-ácido metionina no organismo *Mycobacterium tuberculosis*. Para uma melhor visualização da complexa estrutura de metabolismo que está compactada neste retângulo, incluímos no Apêndice C uma imagem desse caminho metabólico como

é disponibilizada na base de dados KEGG. O objetivo final do método Pathmod é obter escores de atividade e consistência para o caminho metabólico (retângulo amarelo). Essas métricas serão obtidas como o escore médio de todas as reações metabólicas que são parte desse caminho metabólico (retângulos vermelhos). O escore de *atividade* indicará o quão provável é a atividade de um caminho metabólico, enquanto o escore de *consistência* indicará o quão provável é que o resultado do escore de atividade para a reação esteja consistente com os estados de ativação dos genes de entrada (enzimas).

Inicialmente, o método identifica automaticamente todas as reações metabólicas que fazem parte deste caminho metabólico na base de dados KEGG, e busca o código dessas reações no modelo metabólico GSMN-TB. Na rede metabólica usada neste trabalho, as aproximadamente 980 reações metabólicas são indexadas pelos números *R001*, *R002*, etc. Para cada uma das reações metabólicas encontradas, a rede é pesquisada novamente para identificar todas as enzimas que catalisam essas reações. Para este exemplo, o método identifica que no caminho da biossíntese da metionina fazem parte as reações *R243*, *R255* e *R260*, entre outras. Embora o caminho de produção de metionina seja composto de um grande número de reações metabólicas (ver Figura 38), para simplificar o exemplo, mostraremos apenas as três reações metabólicas indicadas acima. Essas reações processam alguns dos metabólitos necessários para a síntese de moléculas de metionina.

Na Figura 27, a reação metabólica *R243* é catalisada pela enzima *metH* (gene *Rv2124c*), a reação metabólica *R260* é catalisada pela enzima *metB* (gene *Rv1079*), e a reação *R255* é catalisada pela ação combinada das duas enzimas *metB* e *metK* (gene *Rv1392*). Sempre que uma reação é catalisada por mais de uma enzima, os seus escores são combinados em uma expressão que representa a atividade catalítica combinada do conjunto das enzimas. Como um exemplo, a reação *R255* é catalisada pelas duas enzimas *rv1392* e *Rv1079*, combinadas na expressão:

$$\texttt{R255} = \texttt{Rv1079} \quad \text{OR } \texttt{Rv1392} \tag{4.11}$$

indicando que essa reação metabólica é catalisada quando uma ou a outra, ou ambas, enzimas estão presentes no ambiente celular[2]. Outras expressões metabólicas podem incluir mais enzimas, e essas são sempre combinadas por dois operadores denominados {OR, AND}. Para calcular as probabilidades a partir dessas expressões utilizamos um método

[2]Embora importante, porém necessário para simplificar o método, assume-se que a atividade catalítica não é dependente de outros fatores externos à rede metabólica tais como a presença de metabólitos ou íons específicos no citoplasma.

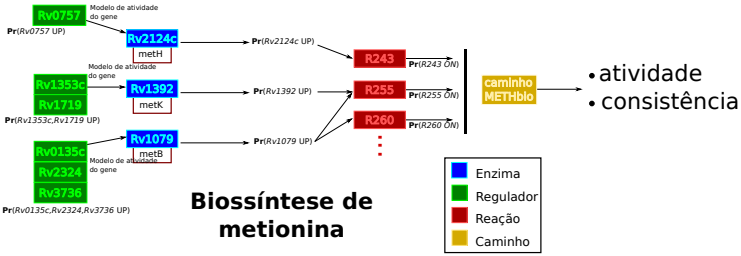

Figura 27 – Representação esquemática do método proposto Pathmod. Exemplo de cálculo dos escores de atividade e consistência para caminhos metabólicos. Nesse exemplo mostramos as proteínas fatores de transcrição (retângulos verdes) para as quais as enzimas (retângulos azuis) que são seus alvos de regulação já são conhecidos. Estas enzimas são responsáveis por catalisar as reações metabólicas (retângulos vermelhos) que participam do caminho metabólico (retângulo amarelo) para biossíntese do amino-ácido metionina no *Mycobacterium tuberculosis*. *Fonte: Adaptado de (EFRONI et al., 2011).*

similar ao apresentado em (COLIJN et al., 2009), substituindo o operador `OR` pela operação max(·) e o operador `AND` pela multiplicação. Esse procedimento garante que todas as probabilidades de reações obtidas das enzimas sejam probabilidades válidas.

Os escores das enzimas são obtidos de forma similar ao método PathOlogist original, como resultado das probabilidades de ativação dos fatores de transcrição (retângulos verdes) que regulam cada uma das enzimas. Essas interações regulatórias podem ser obtidas de diversas maneiras. Uma possibilidade é o uso de redes de regulação publicadas em artigos como (BALÁZSI et al., 2008; SANZ et al., 2011). Em geral estas redes apresentam interações melhor validadas (melhor qualidade), porém apresentam apenas um conjunto reduzido de todas as interações regulatórias disponíveis (menor quantidade). Outra opção é usar diretamente as informações de regulação disponíveis em bases de dados. Nesse caso a quantidade de interações que podem ser obtidas é maior, porém a qualidade é determinada automaticamente e é necessário ter em mente que nem todas apresentam validação biológica experimental. Para este trabalho utilizamos a lista apresentada em (SANZ et al., 2011) aumentada de alguns fatores de transcrição importantes obtidos da lista de genes reguladores do *Tuberculosis Database*, através do formulário online (TB Database, 2015a).

No exemplo da Figura 27, a probabilidade de que o gene *Rv1392c*

esteja em seu estado ativado (UP) é calculada pela probabilidade dos seus dois fatores de transcrição *Rv1353c* e *Rv1719* estarem também em um estado de alta expressão genética na condição sob estudo, enquanto a transcrição do gene *Rv1079* é regulada pelos três FTs *Rv0135*, *Rv2324* e *Rv3736*. As probabilidades de ativação desses fatores de transcrição são obtidas de dados de expressão transcriptômica (mRNA) e do modelo probabilístico para a expressão de cada gene (indicado na Figura 27 como modelo de atividade do gene), da mesma forma como feito no método PathOlogist. Por exemplo, no caso das reações do METHBio, a enzima *metH (Rv2124c)* é positivamente regulada pelo *hub phoP (Rv0757)*; enzima *metK (Rc1392)* é positivamente regulada pelos genes *Rv1353c* e *Rv1719*; etc. De acordo com a base de dados *Tuberculosis Database* (TB Database, 2015b) e os resultados de Sassetti et al (SASSETTI; BOYD; RUBIN, 2003), dessas três enzimas, apenas *metK* é essencial para sobrevivância do *Mycobacterium tuberculosis*.

Escores de consistência são calculados de forma similar, porém ao invés de serem usadas as probabilidades $P(UP|x)$ dos fatores de transcrição de cada enzima, as probabilidades para as enzimas são obtidas diretamente a partir do seu valor de expressão. A concordância entre a atividade indicada pelo conjunto de fatores de transcrição e a atividade enzimática dada pelo valor de expressão obtido diretamente da micromatriz é uma medida de quão bem a rede metabólica e fatores de transcrição consegue predizer a atividade enzimática a partir de dados de mRNA. Essa é uma importante métrica a se ter em modelos *in silico*, pois a partir de uma metodologia iterativa de biologia computacional, descrita conceitualmente em (PALSSON, 2000), escores podem ser calculados, com baixa consistência sendo um indicativo de problemas na associação entre reguladores e regulados ou entre enzimas e reações. As premissas e dados usados no modelo podem ser revistos e a metodologia reaplicada, refinando constantemente o modelo e as previsões.

É fácil ver que o método Pathmod é bastante flexível e permite o estudo mesmo da evolução da atividade metabólica ao longo do tempo, se experimentos de micromatrizes de DNA longitudinais estiverem disponíveis. Esses aspectos, associados a uma facilidade de interpretação dos escores em termos de atividade e consistência, tornam o Pathmod uma boa opção para ser usada por investigadores pouco familiarizados com aspectos mais técnicos das técnicas da biologia computacional.

Embora tenhamos mostrado que a metodologia apresentada é útil no estudo de atividade de caminhos metabólicos, os resultados dependem em grande parte da qualidade das fontes de dados e redes *in*

silico (tanto metabólica quanto regulatória) utilizadas, e à medida que novas redes vão sendo apresentadas e fiquem disponíveis, com melhores associações entre genes e enzimas, e mais validações experimentais confiram maior confiabilidade às informações presentes nas bases de dados, é claro que os resultados das predições e a qualidade dos escores serão cada vez melhores.

4.5 IMPLEMENTAÇÃO

O PathOlogist Modificado para Metabolismo (Pathmod) foi implementado na linguagem de programação R, e todos scripts e conjuntos de dados estão disponíveis para utilização e modificação por solicitação ao autor. A escolha de R para a implementação do método se deu por essa ser uma linguagem funcional de alto nível e qualidade que compartilha características de uma linguagem de *scripting* com outras linguagens, além de ser *open-source* e sem custo, ao contrário de outras linguagens que impõem maiores restrições comerciais. Outra vantagem de R são os muitos pacotes de funções para análise genética e outras tarefas de bioinformática disponíveis através do portal Bioconductor (Fred Hutchinson Cancer Research Center, 2015). A ideia e base para implementação e aplicação do método PathOlogist para metabolismo foi o uso da rede metabólica descrita em (BESTE et al., 2007), a GSMN-TB, uma compilação de alta qualidade de informações metabólicas sobre o *Mycobacterium tuberculosis*, frequentemente utilizada em estudo de metabolismo baseados em FBA, como apresentado no capítulo anterior. Para atividade regulatória utilizamos uma versão aprimorada da rede apresentada em Sanz et al. (SANZ et al., 2011). Uma atualização dessa rede foi necessária pois ainda que a rede Sanz tenha sido curada manualmente, e portanto validada com a literatura relevante, mais recentemente muitas interações genéticas foram melhor esclarecidas e outras novas descobertas, não refletindo portanto a rede original. Scripts de busca e pesquisa em bases de dados desenvolvidos originalmente em Perl foram transcritos para R de forma a melhor integrar as diferentes partes do modelo. Embora não apresentado, para classificar a atividade regulatória dos fatores de transcrição em inibitória ou de ativação, foram utilizados os resultados de uma recente publicação bastante completa sobre os fatores de transcrição do *Mycobacterium tuberculosis*(RUSTAD et al., 2014).

Na próxima seção apresentamos alguns resultados obtidos com o método Pathmod a partir das micromatrizes do conjunto de dados

Boshoff e experimentos com o antibiótico mefloquina. Descreve-se a utilidade dos escores para separar condições experimentais e a possibilidade de inferência sobre mecanismo de ação de compostos antibióticos para tratamento da tuberculose.

4.6 RESULTADOS

Nesta seção são apresentados resultados de validação do método Pathmod, que indicam a sua boa capacidade de agrupamento e classificação de fenótipos com base na atividade metabólica. É o nosso objetivo mostrar como o método pode ser utilizado para estudar fenótipos em bactérias.

A grande vantagem do método Pathmod, o qual faz um mapeamento de dados crus de alta dimensionalidade para dados de menor dimensionalidade e de mais fácil interpretação, é que ele permite deixar de lado a complexidade da resposta transcripcional da célula resultante de um estresse induzido por antibiótico para estudar diretamente os efeitos produzidos no metabolismo. Na busca por novos compostos antibióticos, é de maior vantagem para o investigador que os dados observados sejam de mais fácil e direta interpretação.

Descrições dos tratamentos discutidos nesta seção foram incluídas no Apêndice B e podem ser encontradas no material original em (BOSHOFF et al., 2004).

4.6.1 Atividade Metabólica na Classificação de Fenótipos

Inicialmente, da mesma forma como apresentado em (VARADAN et al., 2012) e (EFRONI; SCHAEFER; BUETOW, 2007), avaliou-se a capacidade dos escores de caminhos metabólicos calculados com o método Pathmod de classificar corretamente entre condições de controle e de tratamento um determinado composto. Em geral, os escores de apenas alguns caminhos metabólicos obtidos com o Pathmod conseguem classificar corretamente as diferentes condições, visto que nem todos os mecanismos metabólicos são alterados de forma semelhante. Como exemplo, avaliamos o tratamento com o composto natural ascididemina (MATSUMOTO et al., 2003). A ascididemina é um produto natural, i.e. um composto orgânico purificado isolado a partir de fontes naturais. Os produtos naturais são o ponto inicial ou inspiração para o desenvolvimento de diversos antibióticos. A ascididemina possui alta ação

citotóxica e grande atividade antimicobateriana. Segundo (BOSHOFF et al., 2004), uma assinatura que identifica o perfil dessa droga é um aumento significativo na expressão de genes com atividade associada ao ferro e biossíntese de micobactina (VOSS et al., 2000), assim como o dipiridil e a deferoxamina, outros compostos que também apresentam comportamento diferencial nos sequestrantes de ferro.

O conjunto de dados Boshoff contém sete experimentos de micromatriz para avaliação da ascididemina (denominado composto #111895), sendo que dois foram realizados com uma concentração de 0.5μg/mL em relação ao controle, dois com dose de 1μg/mL, e três com a maior concentração, igual a 2μg/mL. Note-se que para este composto a concentração inibitória mínima (MIC) foi identificada e igual a 0.2μg/mL. O tratamento foi observado em dois pontos amostrais (6h e 12h), porém na continuidade serão analisadas apenas as condições experimentais realizadas 6 horas após a exposição ao composto. Decidiu-se descartar a única micromatriz do ponto de 12 horas por apresentar comportamento de *outlier* em relação a todas as outras observações. É possível que esse comportamento seja resultante do uso de uma concentração muito alta (5× MIC) associada a um tempo de exposição muito longo (diferente de todas as outras observações) e, portanto, não auxilia nas validações que deseja-se fazer sobre o uso dos escores Pathmod. Tanto a observação de controle quanto do tratamento de 12 horas não foram incluídas na discussão a seguir.

Para identificar os caminhos metabólicos com melhor capacidade de classificação da ascididemina, os escores de atividade obtidos com Pathmod foram classificados com teste-t para diferença entre médias. Os escores de atividade para o caminho metabólico com a melhor separação entre as médias estão mostrados na Figura 28. É significativo que o caminho metabólico representado na figura (*mtu01053*), seja o caminho dos sideroféros (transporte de ferro)[3], visto que estes já foram identificados como a melhor assinatura do perfil deste composto. Observa-se na figura que os escores de atividade para todas as condições de controle (CTL) são consistentemente altos (escores próximos de 0.95), enquanto que os escores do tratamento apresentam, também de forma consistente, escores menores (próximos de 0.80). Observa-se que o escore de atividade é claramente dependente da dosagem.

Outros caminhos metabólicos cujos escores apresentaram as maiores variações entre as condições de controle e tratamento para o com-

[3]Uma listagem completa dos caminhos de acordo com a base de dados KEGG (Kanehisa Labs, 2015) está incluída no Apêndice A.

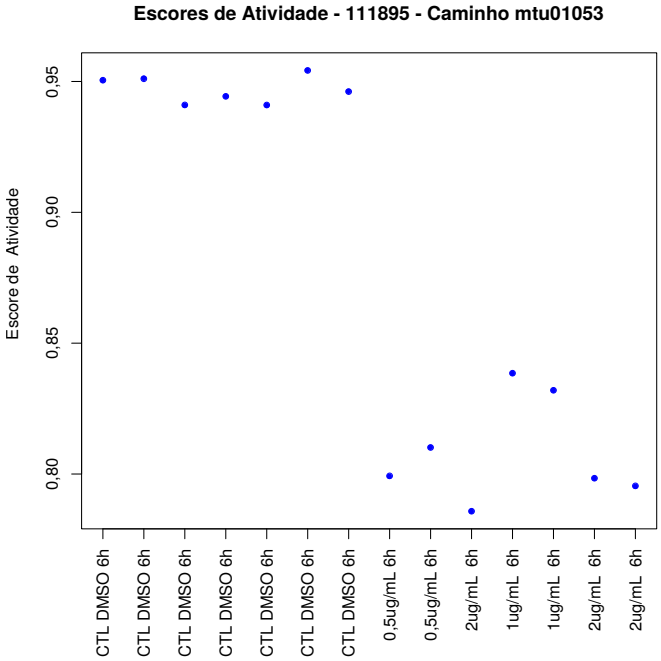

Figura 28 – Escores de atividade para o caminho metabólico *mtu01053* (Biossíntese de Sideróforos e Grupo de Peptídeos Não-ribossômicos) para o tratamento do *Mycobacterium tuberculosis* com o composto #111895 (ascididemina). Observa-se que a atividade deste caminho metabólico é uma assinatura importante para classificar esse tratamento. *Fonte: Elaborada pelo autor.*

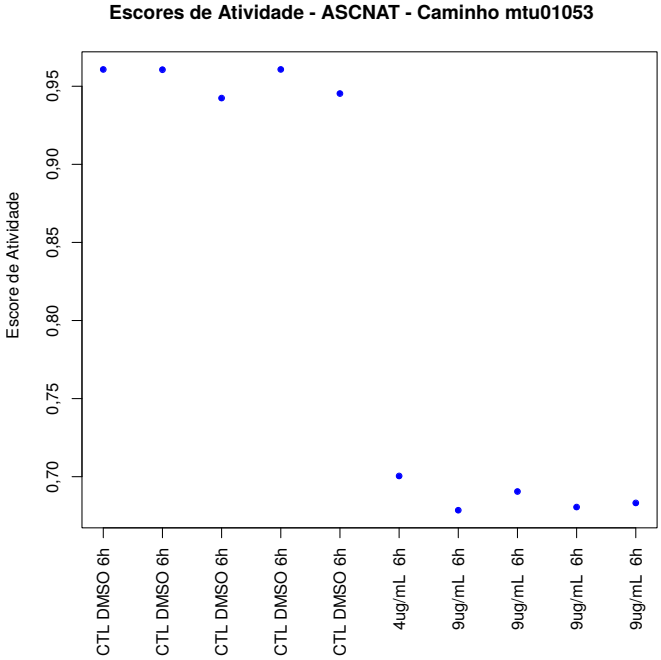

Figura 29 – Escores de atividade para o caminho metabólico *mtu01053* (Biossíntese de Sideróforos e Grupo de Peptídeos Não-ribossômicos) para o tratamento do *Mycobacterium tuberculosis* com o produto natural bruto fonte de ascidedimina. Observa-se que a atividade deste caminho metabólico é uma assinatura importante para classificar esse tratamento e apresenta resposta bastante similar à do composto #111895. *Fonte: Elaborada pelo autor.*

posto ascididemina[4] foram *mtu00061* e *mtu01040* de biossíntese de ácidos graxos e ácidos graxos não-saturados (componentes de micobactinas), *mtu00300* de biossíntese de lisina (envolvida na síntese de micobactinas (FRANKEL; BLANCHARD, 2008) e um possível alvo para drogas anti-tuberculose) além das funções metabólicas de síntese de D-glutamato e D-glutamina, os quais embora ainda não tenham sido relacionados às micobactinas, possivelmente também estão associados à resposta do *Mycobacterium tuberculosis* à ascididemina e seu produtos naturais.

Também foram realizados experimentos com o produto natural bruto (não purificado), fonte da ascididemina, retirado diretamenre do organismo *Eudistoma amplum*, um organismo marinho da classe *Ascidiacea*. Segundo (BOSHOFF et al., 2004), o composto bruto e a ascididemina purificada produziram respostas transcriptômicas *de perfis muito similares*. Com o método Pathmod, os escores de atividade obtidos foram bastante similares para os dois compostos, porém com duas ressalvas interessantes. Alguns caminhos metabólicos para o produto bruto apresentaram significativo aumento na quantidade de ruído biológico para o produto bruto, uma possível consequência do produto não purificado conter diversos outros elementos que podem afetar de maneira diferenciada e não específica o metabolismo do *Mycobacterium tuberculosis*. Apresentamos na Figura 29 os escores de atividade para o mesmo caminho metabólico da Figura 28. Pode-se observar que os dois compostos afetam o caminho metabólico de forma similar (inibição), sendo que os escores menores são obtidos para maiores concentrações do composto (com um aumento não explicado no caso das concentrações de 1μg/mL para a ascididemina). Observa-se que os escores para o produto bruto apresentam uma diferença de 0.13 em relação ao composto purificado.

Ainda que os escores para vários caminhos metabólicos para o produto bruto e a ascididemina tenham sido bastante similares, foram notadas algumas diferenças mais significativas em alguns caminhos. Em relação à resposta do produto natural bruto, o composto #111895 (ascididemina) causou uma menor inibição média no caminho *mtu00061*, maior ativação do caminho *mtu01040* e maior inibição dos caminhos *mtu000480* e *mtu000730*. Ainda para o caminho *mtu000523*, o composto ascididemina apresentou escores menores em relação ao controle, ao passo que o produto natural causou uma maior ativação deste caminho.

Ainda em relação a esses dois compostos, é importante observar

[4]Dados não mostrados.

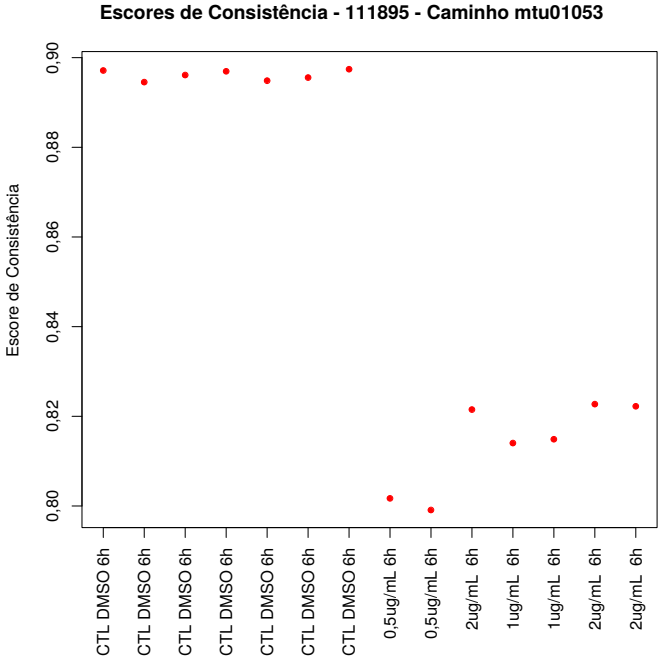

Figura 30 – Escores de consistência para o caminho metabólico *mtu01053* (Biossíntese de Sideróforos e Grupo de Peptídeos Não-ribossômicos) para o tratamento do *Mycobacterium tuberculosis* com o composto #111895 (ascididemina). *Fonte: Elaborada pelo autor.*

a diminuição dos escores de consistência para o composto bruto, uma possível indicação de que este composto produz uma resposta menos específica que o composto purificado. Na Figura 30 são mostrados os escores de consistência para a ascidemina, ainda para o mesmo caminho metabólico *mtu01053*, ao passo que os escores correspondentes para o produto bruto natural estão mostrados na Figura 31.

Um outro resultado importante que pode ser obtido dos escores de atividade obtidos com o modelo Pathmod, é simplesmente a identificação dos níveis de atividade dos caminhos metabólicos. O modelo atual calcula escores de atividade para 82 caminhos metabólicos (alguns com sobreposição), que podem ser usados para identificar quais áreas metabólicas apresentam não apenas uma maior variação entre condição de tratamento e controle, mas a magnitude desses níveis de

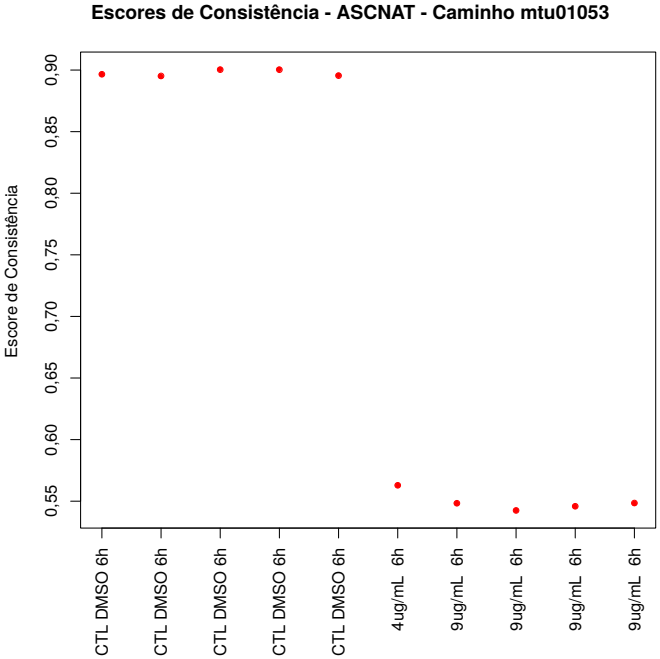

Figura 31 – Escores de atividade para o caminho metabólico *mtu01053* (Biossíntese de Sideroféros e Grupo de Peptídeos Não-ribossômicos) para o tratamento do *Mycobacterium tuberculosis* com o produto natural bruto fonte de ascidedimina. *Fonte: Elaborada pelo autor.*

atividade, o que pode ser útil na caracterização do metabolismo sob uma determinada condição experimental. Por exemplo, no caso anterior do composto natural ascididemina, dos 82 escores de atividade, um caminho apresentou escores médios para o grupo de tratamento na faixa entre 0,4–0,5, ao passo que aproximadamente metade dos caminhos apresentou escores de atividade altos, na faixa de 0,9–1,0. A Tabela 4.6.1 mostra as porcentagens de caminhos metabólicos em cada uma das faixas não-vazias.

Tabela 8 – Porcentagem de caminhos metabólicos com escores em cada uma das faixas de valores indicadas na primeira coluna. A porcentagem é sobre um total de 82 caminhos metabólicos.

Faixa	%	número de caminhos (total 82)
(0,9 – 1,0]	48,8%	40
(0,8 – 0,9]	29,2%	24
(0,7 – 0,8]	15,8%	13
(0,6 – 0,7]	4,9%	4
(0,5 – 0,6]	—	0
(0,4 – 0,5]	1,22%	1

► Esta tabela mostra a quantidade de caminhos metabólicos com escores de atividade médio para a condição de tratamento com ascididemina em cada uma das seis faixas de valores indicadas na primeira coluna. Observa-se que grande parte dos caminhos metabólicos apresentam atividade metabólica mais alta, com apenas 5 caminhos apresentando escores de atividade abaixo de 0,7.

Uma segunda validação importante é verificar a capacidade de classificação de condições experimentais para cada um dos 75 tratamentos individuais. Utilizando um algoritmo de agrupamento hierárquico[5] usando como métrica de dissimilaridade a distância euclidiana entre vetores de escores de atividade, verificou-se que dos 75 tratamentos disponíveis no conjunto de dados Boshoff, o processo de agrupamento para cada tratamento individual classificou corretamente as condições experimentais separando de forma clara controles e tratamentos. Os vetores de escores de atividade metabólica dos tratamentos também foram classificados corretamente por instantes amostrais e concentração do composto ou intensidade do tratamento (e.g. no caso de radiação ul-

[5]Todas análises realizadas na linguagem de programação R. Algoritmo de agrupamento *(clustering)* `hclust()` e função de dissimilaridade `dist()`, ambas no pacote `::stats`

travioleta). Poucas classificações incorretas foram identificadas, como algumas condições com radiação ultravioleta e com o composto PA-1, por exemplo.

Nas Figuras 32 e 33 são apresentados os dendrogramas obtidos com o processo de agrupamento hierárquico para dois tratamentos individuais:

- **deferoxamina**: 4 condições de controle, sendo uma contendo DMSO, e 4 condições de tratamento, sendo três a uma concentração de 150μM e uma com concentração maior igual a 250μM (MIC não indicada).
- **cefalexina**: 7 condições de controle, sendo duas contendo DMSO e duas contendo EtOH. 7 condições de tratamento separadas em duas concentrações: cinco com concentração de 100μg/mL e duas com concentração de 20μg/mL (MIC não indicada)

Para os dois tratamentos apresentados na Figuras 32 e 33, todas as condições experimentais são separadas corretamente pelo algoritmo de agrupamento com base nos vetores de escores de atividade metabólica calculados com o método proposto PathMod. Embora não mostrados, os resultados do agrupamento individual por tratamento são similares para todos os 75 tratamentos analisados em (BOSHOFF et al., 2004) e indicados na Tabela B.1 para referência.

4.6.2 Mecanismos de Ação

Um dos usos possíveis para o Pathmod é o uso dos escores de atividade e consistência para agrupamento do mecanismo de ação de compostos antibióticos. Por exemplo, no caso do *Mycobacterium tuberculosis*, o conhecimento da informação sobre o mecanismo de ação é menos essencial para o desenvolvimento de uma nova droga antituberculose do que para compreender como funciona o mecanismo de desenvolvimento de resistência do organismo a um composto. Visto que o Mtb apresenta alta adaptabilidade e plasticidade regulatória e metabólica, a compreensão dos mecanismos de ação e de resistência são complexos e podem se manter elusivos por décadas, mesmo depois da eficácia terapêutica de um composto ter sido verificada como, por exemplo, no caso da isoniazida (TIMMINS; DERETIC, 2006). Compreender como funciona a ação bactericida de um composto antibiótico para o qual o organismo desenvolve resistência é de primordial importância para definir um plano de ação que permita combater o desenvolvimento

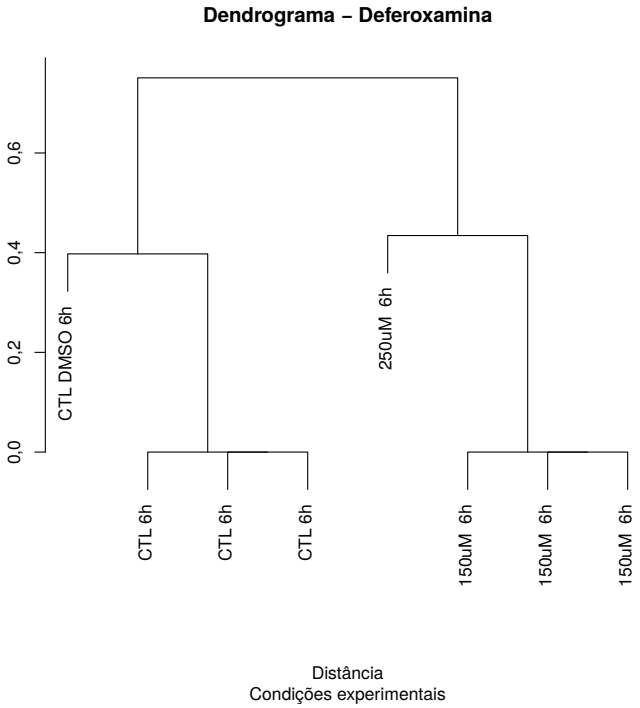

Figura 32 – Dendrograma de todas condições experimentais do tratamento do Mtb com o composto deferoxamina. A árvore foi obtida por agrupamento hierárquico dos escores de atividade metabólica calculados com o método Pathmod. Observa-se que as condições experimentais são corretamente agrupadas por tipo e concentração do composto. *Fonte: Elaborada pelo autor.*

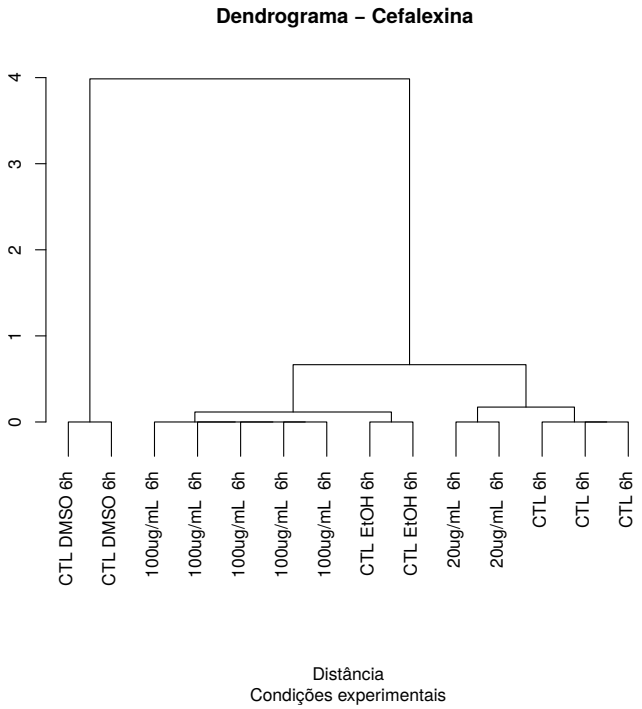

Figura 33 – Dendrograma de todas condições experimentais do tratamento do Mtb com o composto cefalexina. A árvore foi obtida por agrupamento hierárquico dos escores de atividade metabólica calculados com o método Pathmod. Observa-se que as condições experimentais são corretamente agrupadas por tipo e concentração do composto. *Fonte: Elaborada pelo autor.*

das cepas resistentes e a busca por drogas de ação sinergística.

A seguir apresentamos uma breve discussão das assinaturas metabólicas obtidas com o Pathmod para os grupos de compostos e tratamentos estudados no conjunto Boshoff. O intuito desta discussão é mostrar que os escores Pathmod são úteis não apenas para identificar mecanismos de ação similares, mas também para compreender o funcionamento desses mecanismos de ação de forma individual. Os compostos e as assinaturas metabólicas avaliadas a seguir demonstram que os escores do modelo estão correlacionados com vários dos mecanismos de ação já identificados de alguns compostos. Nos casos em que há divergências, há possibilidades de encontrarmos outras vias de ação ainda não identificadas.

Os 75 tratamentos estudados em (BOSHOFF et al., 2004) e listados na Tabela B.1 (Apêndice B), que possuem mecanismo de ação identificado, podem ser classificados de forma geral em oito classes de ação bactericida. Essas classes, segundo (BOSHOFF et al., 2004) são:

- inibição da síntese da parede celular
- piridoacridonas e sequestrantes de ferro
- amidas aromáticas hidrolizadas no ambiente intracelular
- inibição de síntese proteica
- inibição da respiração por agentes diferentes de NO
- crescimento sob condições associadas com a expressão do *regulon* DoS
- agentes causadores de danos à integridade ou topologia do DNA
- inibidores transcripcionais

É importante notar que diversos compostos ou não apresentam um mecanismo de ação totalmente específico (drogas de amplo espectro, espectro dual, etc.) ou apenas uma faceta do seu mecanismo de ação foi identificada até o presente. De forma que, as classes identificadas acima são uma classificação geral que é mais ou menos aparente de acordo com os experimentos realizados com cada um dos compostos. Por exemplo, para o composto que será avaliado na próxima seção e para o qual o mecanismo de ação exato ainda é desconhecido, a mefloquina, mais de uma publicação já identificou uma possível ação não-específica sobre o organismo, na forma de processos de detoxificação e efluxo transmembrana (DANELISHVILI et al., 2005; PIEDADE et al., 2014; MONTEZANO

et al., 2015). O conjunto de dados Boshoff contém experimentos com depleção de nutrientes, crescimento em meios acídicos ou com fontes de carbono diferentes de glicose. Estes tratamentos não foram incluídos na análise a seguir.

Inicialmente, para o estudo dos mecanismos de ação, realizou-se o procedimento de agrupamento usando k-means para agrupar os diferentes tratamentos, com a função R `kmeans()` do pacote `stats`. O procedimento foi realizado 10 vezes com diferentes inicializações para os agrupamentos *(clusters)* e o resultado com agrupamentos mais compactos foi utilizado para análise (selecionado com a métrica `withnss` retornada pela função (THEODORIDIS; KOUTROUMBAS, 2008; R Core Team, 2013)). Entretanto, observou-se que os resultados não diferiram de forma significativa para as diferentes realizações e as conclusões apresentadas em geral se mantêm consistentes.

O conjunto de dados agora inclui as micromatrizes de todos os tratamentos antibióticos. Os resultados estão apresentados na Tabela 4.6.2 e indicam que embora algumas observações tenham sido agrupadas de forma aparentemente incorreta, para a maior parte dos tratamentos o agrupamento ocorreu de acordo com os mecanismos de ação já conhecidos. Como exemplo de um agrupamento correto, verifica-se que os agrupamentos 4, 6 e 10 mostrados na Tabela 4.6.2 compreendem apenas compostos cujo mecanismo de ação é a inibição da respiração (classe 5 do Apêndice B). É possível que os compostos agrupados em cada um destes três agrupamentos (4,6,10) estejam de alguma forma mais relacionados entre si em termos metabólicos, porém essa avaliação não foi realizada. Outro agrupamento correto ocorre com os tratamentos com peróxido de hidrogênio (H_2O_2), raios ultravioleta e MTM (agrupamento 7 e classe 7). Como exemplo de um agrupamento incorreto, observa-se que o mecanismo de ação conhecido e aceito para a Levofloxacina e Ofloxacina são danos causados ao DNA (classe 7 no Apêndice B), porém estes dois tratamentos ficaram agrupados com compostos cujo mecanismo de ação aceito é a inibição da parede celular (agrupamentos 1,5,8 na Tabela 4.6.2. O composto tetraciclina também ficou agrupado incorretamente, pois apesar do seu mecanismo de ação aceito ser a inibição da síntese proteica (classe 4), permaneceu agrupado com outros compostos que inibem a síntese da parede celular (agrupamentos 1,5,8).

Esse resultado sobre a capacidade de agrupamento dos escores Pathmod é muito importante, pois mostra que o mecanismo de ação de muitas drogas pode ser identificado a partir das alterações metabólicas efetuadas pelo *Mycobacterium tuberculosis* após sua exposição.

Tabela 9 – Agrupamento dos tratamentos antibióticos e mecanismos de ação.

Cluster	Mecanismo de ação	Tratamentos
2,3	inibição de síntese proteica	Amikacina, Capreomicina, Roxitromicina, Estreptomicina
2	inibidor transcripcional	Rifampicina, Rifapentina
1,5,8	inibição síntese parede celular	TLM, #109, #241, #59, Ampicilina, DIPED, Cefalexina, Isoniazida, Etionamida, Etambutol, Cerulenina, Antimicina A, Procept 6776, mefloquina, PA-1, PA-21, *Tetraciclina**, *Ofloxacina**, *Levofloxacina**
4,6,10	inibição da respiração	CCCP, DCCD, DNP, DTT, DTNB, $ZnSO_4$,NaN_3, KCN, Methoxatina, Clotrimazol, Cloropromazina, Menadiona, Nigericina, Clofozimina, Econazol, Triclosano, Tioridazina, Verapamil, Valinomicina, PA824, ARP2, ARP4, *Novobiocina**
9	piridoacridinas	AscNatProd, #121940, #124196, #111895, #111891, Deferoxamina, Dipiridil, Procept6778
7	danos ao DNA	H_2O_2, MTM, UV
11	amidas aromáticas	PZA, 5CL-PZA, BZA, Nam

► Esta tabela mostra os resultados do agrupamento dos tratamentos antibióticos do conjunto Boshoff. A primeira coluna indica os agrupamentos, a segunda coluna indica o tipo de mecanismo de ação dos compostos associados aos agrupamentos, e a terceira coluna indica os tratamentos associados a estes agrupamentos. Compostos indicados com um asterisco tiveram mais de 75% das suas condições associadas com agrupamentos diferentes do mecanismo de ação geralmente aceito.

Analisando os agrupamentos obtidos, várias observações podem ser feitas. A primeira observação é que compostos com mecanismos de ação similares foram agrupados corretamente. Entretanto, alguns mecanismos de ação fundiram-se em um único agrupamento, como é

o caso dos inibidores transcripcionais Rifampicina e Rifapentina, que ficaram associados com mecanismo de inibição de síntese proteica. A explicação possível para esse comportamento é que os mecanismos de ação desses compostos não permitem diferenciá-los a partir do metabolismo, mas apenas em um nível anterior, com análise transcripcional (síntese de mRNA) ou traducional (síntese proteica). Além disso, o desenvolvimento de resistência a esses compostos, caso observado, possivelmente não possui ligação com alterações metabólicas. Certamente, essa conclusão deve ser tomada com cautela, visto que o próprio método é incompleto, em que as redes regulatórias e metabólicas utilizadas possuem muitos pontos que são desconhecidos e, quando incluídos em uma versão atualizada das redes, poderiam fornecer a resposta e a separação que o método atual não é capaz de prover.

Uma segunda observação representa o outro lado dessa questão, isto é, mecanismos de ação que ficam espalhados por mais de um agrupamento, como compostos inibidores de respiração (agrupamentos 4,6,10) e inibidores da síntese da parede celular (agrupamentos 1,5,8). Nesse caso, os escores metabólicos do Pathmod fornecem um tipo de informação novo, que poderia ser utilizado como uma sub-classificação dos mecanismos de ação em relação ao seu efeito sobre o metabolismo. Esse detalhamento será apresentado a seguir quando selecionarmos subconjuntos de caminhos como uma assinatura metabólica de cada composto.

No caso do mecanismo de ação sobre a integridade do DNA, i.e. compostos que atuam causando danos ao DNA, degradando-o ou modificando a sua estrutura, vários compostos foram agrupados de forma aparentemente incorreta. Levofloxacina e Ofloxacina foram associados aos agrupamentos dos inibidores da síntese da parede celular, enquanto Novobiocina ficou associada completamente com compostos inibidores da respiração, mais próxima dos compostos CCCP, Menadiona e Valinomicina. Os três tratamentos H_2O_2, radiação ultravioleta e Mitomicina agruparam-se em um agrupamento separado desses outros compostos. Tetraciclina, um composto cujo mecanismo de ação aceito é a inibição da síntese proteica, ficou associado com inibidores da síntese da parede celular.

Em relação a compostos examinados no compêndio Boshoff e sobre os quais não há consenso geral sobre o mecanismo de ação, os escores Pathmod indicam as seguintes associações:

- compostos DTNB, Metoxatina, PA824, ARP2, ARP4 e Verapamil foram agrupados com inibidores da respiração

- compostos Antimicina A, Mefloquina, PA-1, PA-21 e Procept 6776 foram agrupados com inibidores da síntese da parede celular

- composto Procept 6778 foi agrupado diferentemente de acordo com a dosagem: concentrações baixas de 5μg/mL foram associadas com inibidores de síntese proteica, enquanto concentrações maiores de 20μg/mL e 40μg/mL foram associadas com piridoacridinas como #111895, #111891 e produto natural bruto fonte de ascididemina.

Em relação ao comportamento de compostos individuais, observou-se que Estreptomicina teve uma condição experimental associada com inibidores de respiração. Os compostos produto natural bruto e #111895 (ascididemina) agruparam algumas condições com Capreomicina, um inibidor de síntese proteica.

Composto Dipiridil ficou agrupado com piridoacridinas apenas para a dosagem mais baixa de 100μM. As condições com dosagem de 200μM foram agrupadas com inibidores da síntese da parede celular.

Embora Mefloquina tenha sido agrupada com a parede celular, observou-se que algumas condições com tempos de exposição mais longos associaram-se com os tratamentos de radiação ultravioleta, peróxido de hidrogênio e mitomicina, i.e., o grupo de causadores de danos à integridade do DNA.

Ainda que Ofloxacina tenha sido agrupada com inibidores da parede celular, as duas condições de Ofloxacina de 10μg/mL foram agrupadas corretamente com o tratamento UV. As maiores concentrações foram agrupadas com de 10μg e ficaram com o composto Tialoctomicina (TLM).

O composto Cefalexina teve uma associação com piridoacridinas em duas condições de 20μg/mL. As concentrações maiores foram corretamente associadas com o mecanismo de inibição da síntese da parede celular.

O composto com mecanismo não identificado, Procept 6778, teve uma associação maior com inibidores da síntese proteica para as condições de menor concentração (5μg/mL). As dosagens maiores de 20μg/mL e 40μg/mL ficaram associadas com piridoacridinas e sequestrantes de ferro. A conclusão a que se chega é que, embora os tratamentos com mecanismos de ação similares fiquem agrupados corretamente com os escores Pathmod, existem variações que são dependendentes da dosagem e que podem ser identificadas nas funções metabólicas.

São interessantes algumas observações feitas sobre dois compos-

tos que claramente não foram agrupados de acordo com o mecanismo de ação conhecido.

Novobiocina Apresenta um comportamento similar ao ácido nalidíxico (SANZEY, 1979) em que inibe a expressão dos *operons* da maltose, galactose e lactose causando repressão catabólica. Com os resultados do Pathmod observa-se claramente que esses caminhos metabólicos apresentam reduzida atividade, quando o Mtb é exposto a concentrações leves de Novobiocina. A exposição do Mtb à Novobiocina também induz o caminho metabólico de síntese de purinas, como indicado em (BOSHOFF et al., 2004). Outras funções metabólicas significativamente alteradas no caso da Novobiocina são o metabolismo de galactose (*mtu00052*), metabolismo de sacarose e amido (*mtu00500*), metabolismo de amino açúcar e açúcar de nucleotídeo (*mtu00520*), inibição da síntese de purinas (*mtu00230*), inibição da biossíntese de LPS[6] (*mtu00540*) e biossíntese da unidade de açúcar de policetídeos (*mtu00521*). É importante notar que, segundo (BOSHOFF et al., 2004), a Novobiocina não é agrupada com outros tratamentos conhecidos que danificam o DNA como a mitomicina, radiação ultravioleta e peróxido de hidrogênio, da mesma forma como obtido com os escores Pathmod.

Levofloxacina Embora apresente o mesmo mecanismo geral de ação de danos à integridade do DNA, é interessante notar que esse composto é considerado um composto de amplo espectro de ação. Os resultados obtidos com Pathmod para a Levofloxacin indicam, de forma semelhante à Novobiocina, que há inibição dos caminhos metabólicos de purinas (*mtu00230*) e pirimidinas (*mtu00240*), indicando o efeito sobre a estrutura do DNA. Porém, o composto apresenta significativa inibição de caminhos associados à síntese de ácidos graxos (*mtu00061, mtu00780, mtu01040*), indicando possivelmente um outro aspecto desconhecido da ação desse composto sobre o Mtb.

4.6.3 Identificação de Descritores Discriminantes

Com esta seção finalizamos a análise dos resultados do Pathmod aplicado ao conjunto de dados Boshoff. O objetivo desta seção é identificar para cada mecanismo de ação um conjunto de caminhos metabólicos cujos escores de atividade classifiquem e separem as condições de tratamento das condições de controle. Esse resultado permite identificar,

[6] LPS (lipopolissacarídeo)

para cada mecanismo de ação, quais caminhos metabólicos produzem a melhor assinatura da ação do composto. As informações a seguir são ilustrativas da capacidade do método Pathmod de sugerir vias de pesquisa para o desenvolvimento de compostos antibióticos, compreensão de mecanismos de ação e de resistência. Tais inferências podem ser utilizadas para direcionar experimentos de bancada que poderão eventualmente auxiliar na identificação de novos compostos anti-tuberculares.

Para cada tratamento foi realizado o procedimento de análise discriminante linear utilizando a função `ldaHmat()` do pacote R `subselect`[7] sobre a matriz de escores de atividade calculados com Pathmod. A seguir, um conjunto de caminhos metabólicos de cardinalidade 5 a 10 que classificasse corretamente as condições experimentais foi obtido com a função `improve()`. Para a seleção dos subconjuntos de variáveis, utilizou-se a matriz de covariância dos escores de atividade para cálculo das matrizes de espalhamento (THEODORIDIS; KOUTROUMBAS, 2008), as quais necessariamente devem ser calculadas com o método de *shrinkage* (SCHÄFER et al., 2015) avaliado no segundo capítulo deste trabalho. O método `improve()` é relativamente sensível a problemas de condicionamento da matriz. Não foi possível executar o método com a utilização de uma estimativa de máxima verossimilhança para a matriz de covariância, mesmo com um alto valor de tolerância (o que produziria resultados cada vez menos acurados). As melhores características apresentadas pela matriz estimada por encolhimento foram essenciais para que pudesse ser utilizado o método `improve()` na seleção dos subconjuntos de escores metabólicos para classificação. Este conjuntos, com os caminhos metabólicos mais significativos para classificação de cada mecanismo de ação, estão indicados a seguir.

Causadores de danos ao DNA Para os compostos Mitomicina, peróxido de hidrogênio e radiação ultravioleta, os caminhos *mtu00053, mtu00230, mu00240, mtu03430, mtu00450, mtu00473, mtu00540, mtu00550, mtu00780, mtu00362* classificam controles e tratamentos com 90% de acerto. Estes caminhos estão associados com a síntese de lipolissacarídeos, biossíntese de ácidos graxos, metabolismo de pirimidinas e purinas, e reparação de bases mal pareadas.

[7] Pacote `subselect`: *Uma coleção de funções as quais (i) avaliam a qualidade de subconjuntos de variáveis como substitutos para um conjunto de dados completo, tanto em análise exploratória de dados quanto no contexto de um modelo linear multivariável, e (ii) buscam subconjuntos que são ótimos segundo uma variedade de critérios.* (CERDEIRA et al., 2015)

Inibidores da Síntese da Parede Celular Para os compostos com mecanismo de ação atuante sobre a composição da parede celular, o subconjunto de caminhos que permite uma classificação com acerto de 90% inclui os caminhos *mtu00051, mtu00061, mtu00072, mtu00561, mtu00473, mtu00780, mtu00565, mtu01040, mtu00540, mtu00550, mtu00790*, i.e. diversos caminhos associados com a produção de lipídeos e ácidos graxos, folato, biotina, peptidoglicanos, D-alanina e degradação de corpos cetônicos.

Inibidores de respiração Para os compostos inibidores de respiração celular, o subconjunto de caminhos metabólicos mais discriminantes inclui *mtu00020, mtu00620, mtu00660, mtu00910, mtu00281, mtu00623, mtu00624, mtu00625, mtu00400, mtu00401.* Caminhos associados com o ciclo tricarboxílico, metabolismo de piruvato, degradação de geraniol, degradação do naftaleno, degradação do tolueno, degradação de cloroalcano e coloralceno, biossíntese de novobiocina e biossíntese de aminoácidos fenilalanina, tirosina e triptofano.

Amidas Aromática *mtu00020, mtu00053, mtu00190, mtu1220, mtu00561, mtu00624, mtu00626, mtu00623, mtu00401, mtu00400, mtu00625* (ciclo do ácido tricarboxílico, metabolismo de ascorbato e aldarato, fosforilação oxidativa, metabolismo de glicerolipídeos, degradação de compostos aromáticos e degradação de hidrocarbonetos policíclicos aromáticos.

Inibidores da Síntese Protéica e Inibidores Transcripcionais Nesta classe, o subconjunto de caminho que permite classificação das condições experimentais com 90% de acerto inclui *mtu00920, mtu00270, mtu00450, mtu00460, mtu00750, mtu00760, mtu00790, mtu2010, mtu03018* (metabolismo de enxofre, metabolismo de cisteína e metionina, metabolismo de compostos selênicos e metabolismo de cianoaminoácido, metabolismo de vitamina B6, metabolismo de ácido nicotínico e nicotinamida, biossíntese de folato, transportadores ABC, degradação de RNA).

Piridoacridinas e Sequestrantes de Ferro Nesta última classe, o subconjunto de caminhos metabólicos cujos escores classificam as condições de tratamento e controle com acerto de 90% inclui *mtu00061, mtu00362, mtu00660, mtu00770, mtu00910, mtu00920, mtu1053* (biossíntese de ácidos graxos, degradação de benzoato, metabolismo de

ácido dibásico derivação 5C, metabolismo de pantotenato e coenzima-A, metabolismo de enxofre e nitrogênio, biossíntese de sideróforos e peptídeos não-ribossômicos).

4.6.4 Mecanismo de Ação da Mefloquina

Anteriormente analisamos micromatrizes de DNA de exposição do *Mycobacterium tuberculosis* à mefloquina. Agora iremos apresentar assinatura metabólica do composto mefloquina que melhor separa as condições experimentais de controle e tratamento.

A mefloquina é um composto da família das quinolonas (FOLEY; TILLEY, 1997) indicado para o tratamento da malária, e apesar dos seus sérios efeitos colaterais psicotrópicos (WEINKE et al., 1991) e do crescente desenvolvimento de resistência por parte do organismo *Plasmodium falciparum* ainda, é um dos tratamentos mais indicados para profilaxia da doença. A mefloquina ainda não possui mecanismo de ação identificado para muitos organismos, incluindo o *Mycobacterium tuberculosis*. Para esse organismo, conforme (DANELISHVILI et al., 2005), seu mecanismo de resistência poderia estar associado com processos de detoxificação e efluxo através da membrana e parede celular, enquanto seu mecanismo de ação poderia indicar uma ação de amplo espectro não-específica. É possível que a mefloquina atue sobre lipoproteínas que interagem com o organismo em caminhos metabólicos utilizados para absorção de fosfolipídeos, no caso do parasita *P. falciparum* (FOLEY; TILLEY, 1997).

Na seção anterior, o procedimento de agrupamento com os outros diversos compostos do conjunto de dados Boshoff agrupou a mefloquina com outros agentes inibidores da síntese da parede celular, embora tenha sido observada uma proximidade com agentes desintegrantes de DNA como a mitomicina e radiação ultravioleta.

Utilizando o procedimento da seção anterior para encontrar um subconjunto de caminhos metabólicos cujos escores representem uma assinatura metabólica para classificar as condições de tratamento e controle, observou-se que os caminhos *mtu00051, mtu00053, mtu00061, mtu00072, mtu00430, mtu00450, mtu00473, mtu00540, mtu00550, mtu00561, mtu00770, mtu00780 e mtu01040* são um conjunto de cardinalidade 12 que permite classificar corretamente as condições experimentais.

Observa-se nesse conjunto que vários caminhos identificados na assinatura da mefloquina são similares aos caminhos encontrados para

outros tratamentos inibidores da síntese da parede celular. Alguns destes caminhos também estão presentes na assinatura de um número de compostos que são agentes causadores de danos ao DNA. Identificam-se também alguns caminhos de outros mecanismos de ação como as piridoacridinas (*mtu00061, mtu00770*), amidas aromáticas (*mtu00053, mtu00561*) e inibição da síntese proteica (*mtu00450*).

Embora esses resultados não esclareçam o mecanismo de ação da mefloquina sobre o *Mycobacterium tuberculosis*, permitem uma apreciação de que a mefloquina, ainda que em termos metabólicos, aja de forma semelhante a outros compostos que inibem a síntese da parede celular ou compartilhe algum mecanismo cujo efeito metabólico seja similar a agentes que causam desintegração ou modificação da estrutura do DNA como mitomicina, peróxido de hidrogênio ou mesmo radiação ultravioleta.

4.7 CONCLUSÕES

Neste capítulo foi apresentada uma nova proposta de estudo do metabolismo do *Mycobacterium tuberculosis* sob diferentes condições experimentais. O método denominado de Pathmod, é baseado no método PathOlogist desenvolvido originalmente para o estudo de fenótipos de câncer, o qual foi aplicado com sucesso na classificação de tumores, e pode ser utilizado para acelerar a descoberta de biomarcadores do câncer.

A adaptação do método é promissora como ferramenta de auxílio no estudo do metabolismo e em áreas mais específicas de interesse para a comunidade científica e clínica, como a descoberta de assinaturas metabólicas de condições experimentais, obtenção de uma melhor compreensão dos mecanismos de ação de compostos antibióticos e mesmo a identificação de mecanismos de resistência desenvolvidos por agentes patogênicos como o *Mycobacterium tuberculosis*.

Embora o trabalho desenvolvido neste capítulo tenha usado o organismo Mtb como ator principal e a associação de mecanismos de ação como motivação, o método Pathmod é suficientemente geral, podendo ser aplicado a qualquer organismo, células eucarióticas ou organismos procariotas, para os quais estejam disponíveis informações genômicas sobre fatores de transcrição e enzimas que possam ser usados para criar uma rede de regulação e uma rede metabólica *ìn silico*, caso ainda não esteja disponível.

O método Pathmod, em sua essência, mapeia dados de expressão

transcriptômica para escores de atividade e consistência metabólica. O mapeamento é vantajoso em dois aspectos: a dimensionalidade do conjunto de dados é reduzida de maneira significativa e, em segundo lugar, os escores, calculados para cada caminho metabólico apresentam um novo aspecto do organismo e são de mais fácil interpretação para o investigador que busca por exemplo, entender as variações de atividade metabólica resultantes da exposição do organismo a um composto antibiótico.

Após a apresentação do método mostrou-se que os escores de atividade metabólica classificam corretamente, para cada tratamento individual, as diferentes condições experimentais (instantes amostrais e concentrações e controles versus tratamentos). Essa classificação foi realizada para diversos tratamentos do conjunto de dados Boshoff. Foram identificadas associações entre escores com maior variação e mecanismos de ação conhecidos para tratamentos individuais. Os escores foram também utilizados para agrupar as condições de tratamento de diversos antibióticos com mecanismos de ação conhecidos. Os agrupamentos obtidos refletem os diversos mecanismos de ação já identificados para os compostos e trazem à tona algumas divergências específicas decorrentes do uso de métricas metabólicas no agrupamento. O processo de agrupamento também permitiu associar alguns compostos com mecanismo de ação não identificado em Boshoff et al. (BOSHOFF et al., 2004).

Ainda foram identificadas assinaturas metabólicas, i.e. subconjuntos de caminhos metabólicos que permitem classificar os compostos de determinado grupo com mecanismo de ação similar. Embora não analisados em detalhes, os resultados indicam uma forte associação entre os caminhos metabólicos identificados e os mecanismos metabólicos de ação dos antibióticos. Finalmente foi apresentado o conjunto de caminhos metabólicos que classifica o composto mefloquina, para o qual um mecanismo de ação ainda não foi identificado. Verificou-se a partir das métricas Pathmod que a mefloquina compartilha uma assinatura metabólica com agentes de inibição da síntese da parede celular e, possivelmente, compostos que causam danos à integridade do DNA, como radiação ultravioleta, mitomicina e H_2O_2.

Aparentemente, o Pathmod é um método novo e promissor para auxílio em estudos do metabolismo. A versão apresentada neste capítulo é apenas um protótipo que demonstra a viabilidade e algumas das possibilidades do método, que entretanto necessita ser aprimorado. Acreditamos que, possivelmente, com uma avaliação mais minuciosa do algoritmo e o desenvolvimento de uma interface de usuário intuitiva e simples, seja possível apresentar à comunidade científica um software

com bom potencial de penetração, e capacidade de se tornar uma ferramenta popular para estudos do metabolismo.

O método será beneficiado se forem realizados procedimentos de validação mais criteriosos, preferencialmente com validação experimental. Estudos que utilizem o método para estudo de metabolismo em escalas menores (apenas determinados conjuntos de caminhos metabólicos e não um estudo em escala genômica) serão certamente bem-vindos. Realizar estudos comparativos com outras propostas de estudo de metabolismo como FBA também são úteis, e permitiriam uma visão diferenciada, centrada na ideia do caminho metabólico, e não mais em genes, proteínas ou reações individuais.

5 MODELO ESTATÍSTICO DO METABOLISMO

5.1 INTRODUÇÃO

No capítulo anterior tratamos especificamente de um problema de classificação, utilizando os escores metabólicos obtidos com o método Pathmod para separar condições experimentais e tratamentos. No presente capítulo trataremos do problema de predição de fluxos metabólicos e da identificação de possíveis caminhos metabólicos de sobrevivência, como na discussão sobre a Figura 20. Apresentamos inicialmente um modelo estatístico bayesiano para modelagem de configurações de fluxo metabólico do *Mycobacterium tuberculosis*. Estuda-se a modelagem com estimação de parâmetros por métodos MCMC (*Markov Chain Monte Carlo)*. Veremos como esse modelo pode ser utilizado para predizer uma configuração metabólica a partir de novos dados experimentais de transcriptômica e a identificação de reações metabólicas (e consequentemente caminhos metabólicos) associadas com metabolismo de sobrevivência após exposição a um determinado composto antibiótico (mefloquina). O uso de conhecimento *a priori* sobre a variação dos fluxos é avaliado.

Este capítulo integra os resultados dos capítulos anteriores em que utilizamos dados de transcriptômica sintéticos gerados a partir da matriz de covariância estimada pelo método de encolhimento (*shrinkage*) apresentado no capítulo 2 e métodos FBA do capítulo 3 para geração de dados de fluxo metabólico. Ambos estes conjuntos de dados são usados na estimação e avaliação do modelo. Os escores metabólicos Pathmod podem, por sua vez, ser utilizados como ferramenta ancilar na detecção de atividade metabólica em situação de sobrevivência.

Aqui, deseja-se estimar um modelo para estudo do metabolismo utilizando dados experimentais de transcriptômica, lembrando que os dois maiores problemas nesse tipo de aplicação são a alta dimensionalidade do conjunto de dados (tanto dos preditores quanto dos dados de saída) e o baixo número de amostras disponíveis para treinamento do modelo.

No caso do modelo apresentado, deseja-se utilizar dados de experimentos de expressão genética, por exemplo, experimentos com micromatrizes de DNA, que produzem valores para os níveis de mRNA de todos os genes do organismo, para predizer configurações metabólicas na forma de fluxos metabólicos. A justificativa e o interesse para a utilização de dados de mRNA como variáveis de entrada do modelo

decorre principalmente da ubiquidade e relativa facilidade com que estes experimentos podem ser realizados (BAR-JOSEPH, 2004; WANG et al., 2008). O uso de dados de fluxos metabólicos como informação de saída decorre do fato que esses são os dados que definem a configuração metabólica do organismo. Dessa forma, o conjunto de dados que utilizaremos na estimação do modelo pode ser indicado como:

$$\mathcal{D} = \{\boldsymbol{e}_i, \boldsymbol{v}_i\} \text{ para } i = 1 \dots n \tag{5.1}$$

em que $\boldsymbol{e}$ indica um conjunto de n amostras de dados multivariados de expressão genética, e $\boldsymbol{v}$ indica um conjunto de dados multivariados de fluxos metabólicos que definem configurações metabólicas correpondentes de saída. No caso do organismo *Mycobacterium tuberculosis*, cada vetor de entrada $\boldsymbol{e}_i$ possui alta dimensionalidade, e é composto por valores de expressão genética para cada um dos 3924 genes atualmente identificados e incluídos nas lâminas de micromatriz utilizadas para os experimentos (Affymetrix[1]). A dimensão de cada vetor de entrada será indicada pela variável p, e portanto para os dados de entrada temos dimensionalidade $p = 3924$.

No caso dos dados de saída, cada $\boldsymbol{v}_i$ inclui os fluxos metabólicos medidos no organismo para a mesma condição experimental em que foram gerados os níveis de mRNA. A dimensão desse vetor depende certamente de quais medidas de fluxos metabólicos são realizadas. Entretanto, experimentalmente, medidas de fluxos metabólicos não são simples de realizar. Em geral, apenas um pequeno número de fluxos pode ser medido, e ainda assim de forma indireta. Por esse motivo, para o estudo do modelo metabólico apresentado neste capítulo, o conjunto de dados de saída de fluxos metabólicos será estimado com o método de análise de balanço de fluxos apresentado no Capítulo 3 deste trabalho. Para cada experimento de micromatriz realizado um vetor de fluxos metabólicos será calculado com a técnica apresentada no Capítulo 3. Esses vetores possuem dimensão igual ao número de reações metabólicas incluídas na rede *in silico* GSMN-TB (BESTE et al., 2007) utilizada no procedimento de FBA, que para esse modelo perfazem 849 reações individuais. A dimensão de cada vetor $\boldsymbol{v}_i$ de saída será indicada pela letra q. Portanto, para o nosso modelo, $q = 849$.

Outra característica do conjunto de dados da expressão (5.1) diz respeito ao número n de observações disponíveis. Embora experimentos com micromatrizes de DNA sejam relativamente fáceis de realizar de

[1] Dados de expressão genética para este trabalho foram obtidos com micromatrizes Affymetrix e processados com os pacotes R Bioconductor `affy` (GAUTIER et al., 2004) e `makecdfenv`.

forma automática, o seu custo ainda é alto para que diversas repetições do experimento sejam realizadas. É de se esperar que o biólogo investigador tenha limitações de custo em seus projetos e que apenas uma quantidade limitada de experimentos possa ser realizada. Dessa forma, é possível que haja mais interesse por parte do investigador em realizar menos repetições de cada experimento e optar pela possibilidade de testar um maior número de condições experimentais, sacrificando a qualidade estatística dos dados. Por exemplo, são comuns experimentos longitudinais em que apenas três ou cinco repetições de cada ponto amostral sejam executadas.

Conclui-se que o conjunto de dados disponível para um estudo do metabolismo apresenta duas características importantes e difíceis de tratar conjuntamente: (1) alta dimensionalidade p dos dados de entrada e pequeno número n de observações. Como já discutido anteriormente, esse tipo de problema de estimação é comumente conhecido por problema "n-pequeno, p-grande". Tais modelos em geral apresentam problemas de identificabilidade, e sofrem significativa degradação nas suas estimativas caso técnicas padrão de estimação sejam utilizadas.

O estudo do metabolismo a partir de medidas de transcriptômica ainda apresenta mais algumas características que são importantes discutir aqui, pois o conjunto de tais condições foi a motivação que levou à escolha do tipo de modelo que será apresentado na próxima seção.

Relação entrada-saída não-linear. O objetivo modelo de predição deste capítulo é modelar a relação que existe entre os dados de expressão genética e fluxos metabólicos em um organismo celular. Entre essas duas representações do funcionamento celular existem diversos processos altamente não-lineares, tais como regulação, tradução, pós-tradução, disponibilidade de nutrientes, alterações de conformação protéica, ação enzimática, entre outros. Todos esses processos tornam a relação entre $\boldsymbol{e}$ e $\boldsymbol{v}$ de difícil estimação com modelos mais simples. Por esse motivo é que a escolha recaiu sobre o uso de um modelo baseado em *kernels* (SCHÖLKOPF; SMOLA, 2002), que permite modelagem de funções entrada-saída não-lineares com uma flexibilidade relativamente maior. A premissa básica para o uso de kernels é que uma relação não-linear no espaço de dados original, pode ser modelada de forma linear em um espaço mapeado.

Número de parâmetros. Uma outra dificuldade, associada com modelos computacionais de sistemas biológicos, é a grande quantidade de parâmetros a estimar. A complexidade do modelo, em termos do

número de parâmetros, é um problema grave quando o número de amostras é pequeno, pois o poder preditivo fica comprometido. O uso de um modelo de regressão linear baseado em *kernels* permite uma redução significativa na quantidade de parâmetros a estimar. Mais especificamente, a escolha de um modelo do tipo máquina de vetores de relevância (BISHOP, 2007) (*relevance vector machine*)) é ainda mais interessante, visto retornar modelos mais esparsos, em que vários parâmetros que não contribuem para a capacidade preditiva do modelo são automaticamente eliminados do mesmo.

Conhecimento *a priori*. Um terceiro e último aspecto importante é o uso de um modelo bayesiano para modelagem da relação $\boldsymbol{e}$ — $\boldsymbol{v}$. Dois aspectos importantes da técnica bayesiana são o uso de distribuições *a priori* para parâmetros do modelo, e a produção não de uma estimativa pontual, como é o caso da estimaçao de máxima verossimilhança, mas de uma distribuição de probabilidade para cada um dos parâmetros do modelo e uma distribuição preditiva das variáveis de interesse para conjuntos de dados de entrada que não tenham sido empregados no treinamento do modelo. Para o caso específico da modelagem do metabolismo do *Mycobacterium tuberculosis*, é nosso intuito mostrar que é possível diminuir a variância das predições de valores de fluxo com o uso de informação *a priori*. Além disso, mostraremos como as distribuições preditivas *a posteriori* podem ser utilizadas para identificação de caminhos de sobrevivência.

5.2 APRESENTAÇÃO DO MODELO

A escolha do tipo de modelo para predição de metabolismo a partir de dados de transcriptômica baseou-se nas dificuldades indicadas na seção anterior: problema n-pequeno, p-grande, com relação não-linear entre dados de entrada e de saída, a busca de um modelo esparso, visto o número reduzido de amostras para treinamento, e a possibilidade de inclusão de conhecimento sobre o problema na forma de distribuiçoes *a priori*.

O modelo RVM (BISHOP, 2007) é uma técnica útil para problemas de regressão, similar às máquinas de vetor de suporte utilizadas em problemas de classificação. A RVM, por apresentar uma formulação bayesiana, permite evitar algumas limitações das máquinas de vetor de suporte como a obtenção de estimativas pontuais (sem caráter probabilístico) e a necessidade de cálculo de parâmetros de complexidade. De

forma simples, a formulação matemática da RVM é um modelo linear de regressão da forma:

$$v(\boldsymbol{e}) = \sum_{i=1}^{n} \beta_i k(\boldsymbol{e}, \boldsymbol{e}_i) + \beta_0 \tag{5.2}$$

em que β_i são os coeficientes de regressão do modelo e as variáveis de entrada não são mais os elementos originais do vetor $\boldsymbol{e}$, mas *kernels* entre estes vetores do conjunto de treinamento. Observa-se que a complexidade do modelo está agora diretamente associada com o tamanho da amostra de treinamento, e portanto um número menor de vetores de treinamento é de alguma forma vantajoso aqui, visto diminuir o número de parâmetros a estimar e, consequentemente, a complexidade do modelo. Na expressão do modelo estatístico RVM da Eq.(5.2), as variáveis preditoras do modelo, i.e. os *kernels*, são produtos internos dos vetores de expressão genética no chamado *espaço das características* (SCHÖLKOPF; SMOLA, 2002). O espaço das características é o espaço para o qual os vetores de entrada são mapeados, através de uma função não-linear $\Phi(\cdot)$. O produto interno de dois vetores nesse novo espaço (com dimensão em geral diferente da dimensão do espaço de dados original), pode ser calculado usando a função *kernel* que define o espaço. No modelo RVM, são as variáveis preditoras do modelo de regressão linear. Uma representação gráfica do mapeamento dos vetores de entrada de expressão genética para o espaço das características e a predição de uma configuração de fluxos metabólicos $\boldsymbol{v}$ com RVM está apresentada na Figura 34. Uma desvantagem do uso de preditores na forma de kernels é que não é mais possível dar uma interpretação biológica para os preditores, como é comumente feito com modelos mais simples de regressão linear e modelos generalizados (GELMAN; HILL, 2007), visto o mapeamento da RVM ser altamente não-linear.

Note-se que embora apenas um fluxo metabólico (univariado) esteja sendo mostrado na Eq.(5.2), o modelo deste capítulo é multivariado, produzindo como estimativa um vetor $\boldsymbol{v}$ de saída.

O modelo RVM é um método bayesiano hierárquico (BISHOP, 2007), em que existem parâmetros e hiperparâmetros a estimar. Por ser um método bayesiano, para cada parâmetro do modelo e para cada hiperparâmetro, é necessário definir distribuições *a priori*. Inicialmente, e sempre que possível, a escolha recai sobre distribuições conhecidas como *distribuições conjugadas*, que permitem maior tratabilidade matemática, permitindo às vezes a solução analítica do problema. Entretanto, em casos de modelos complexos, nem sempre a facilidade do

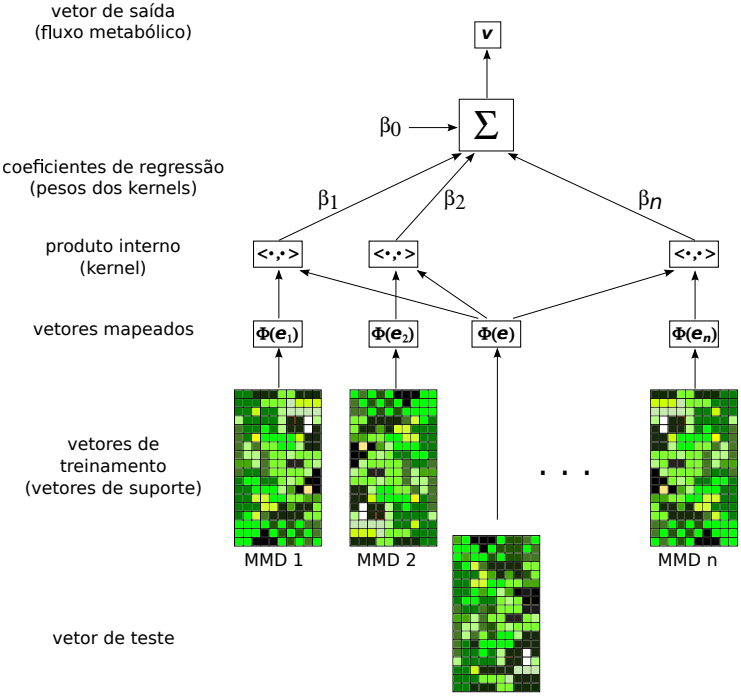

Figura 34 – Representação esquemática do modelo de máquina de vetor de relevância para o problema de regressão entre vetores de expressão genética $\boldsymbol{e}$ e configurações de fluxos metabólicos $\boldsymbol{v}$. Os vetores de entrada $\boldsymbol{e}_i$ do conjunto de dados de treinamento são mapeados para o espaço das características com a função $\Phi(\cdot)$ (a qual não necessita ser conhecida). Após o treinamento da máquina, um vetor de teste $\boldsymbol{e}$ é apresentado ao modelo, o qual produz uma estimativa de fluxo metabólico como uma combinação linear dos produtos internos do vetor de teste com cada um dos vetores de treinamento $\boldsymbol{e}_i$ considerados relevantes durante o treinamento. Os vetores $\boldsymbol{e}_i$ que não contribuem de forma significativa para a qualidade das estimativas durante o treinamento são efetivamente eliminados do modelo, visto que os coeficientes β_i correspondentes a tais vetores são levados a zero durante o treinamento pelo mecanismo de relevância do método. *Fonte: Adaptado de (SCHÖLKOPF; SMOLA, 2002).*

tratamento matemático pode ser garantido, e em tais casos os procedimentos de estimação dos modelos recaem sobre técnicas de amostragem, como usaremos neste capítulo.

O modelo RVM multivariado foi apresentado e avaliado em (CHAKRABORTY; GHOSH; MALLICK, 2012), e é a base do modelo utilizado neste capítulo. O detalhamento matemático, com a derivação das expressões das funções de probabilidade condicionais necessárias para estimação do modelo com MCMC, foi incluído no Apêndice B. Embora em (CHAKRABORTY; GHOSH; MALLICK, 2012) sejam apresentados e comparados diversos modelos para regressão em situações tipo n-grande, p-pequeno, apenas trabalharemos aqui com o RVM multivariável, visto esse ter sido o modelo que apresentou o menor erro de predição para o estudo de simulação com espectroscopia NIR *(near-infrared)*. Um ponto interessante desse modelo é que o parâmetro do kernel utilizado também é estimado juntamente com os outros parâmetros do modelo. Diversos tipos de *kernels* podem ser utilizados, e a escolha do kernel mais apropriado é uma questão difícil de responder *a priori*. O modelo deste capítulo foi estimado com dois tipos de kernel: o kernel gaussiano e o kernel polinomial, apresentados respectivamente na Eq. (5.4) e Eq. (5.3). O *kernel* polinomial aplicado a dois vetores $\boldsymbol{e}$ e $\boldsymbol{e}'$ é dado pela expressão:

$$k(\boldsymbol{e}, \boldsymbol{e}') = (\boldsymbol{e}^T \boldsymbol{e}' + b)^m \tag{5.3}$$

para $b \geq 0$. O kernel polinomial resulta em uma soma ponderada de termos constantes, lineares e todos termos até a ordem m obtidos com os elementos dos vetores passados como argumentos do kernel. No caso específico de $b = 0$ o kernel contém apenas os termos de ordem m. A dimensão do espaço das características é maior que a dimensão original dos dados de entrada, porém essa dimensionalidade mais alta está implícita no kernel e não causa nenhum aumento de complexidade computacional. Por exemplo, no caso de vetores de entrada de dimensão $p = 2$ (vetor $\boldsymbol{e}$ com dois elementos), o uso do kernel quadrático[2] produz um mapeamento dos vetores de entrada para um espaço de dimensão $p_\Phi = 3$.

No caso do kernel gaussiano, a expressão para o kernel é dada por:

$$k(\boldsymbol{e}, \boldsymbol{e}') = \exp\left(\frac{-||\boldsymbol{e} - \boldsymbol{e}'||^2}{2\sigma^2}\right) \tag{5.4}$$

em que σ é o parâmetro do *kernel*. Para definirmos o modelo de dados

[2]kernel polinomial para $m = 2$

apropriado para o problema da estimação do metabolismo, podemos escrever a Eq.(5.2) como:

$$v_i = f(\boldsymbol{e}_i) + \eta_i = \boldsymbol{K}_i^T \boldsymbol{\beta} + \eta_i \tag{5.5}$$

em que $\eta_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. O vetor de coeficientes de regressão $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_n)^T$ é multiplicado pelo vetor

$$\boldsymbol{K}_i = (1, k(\boldsymbol{e}_i, \boldsymbol{e}_1|\boldsymbol{\theta}), \ldots, k(\boldsymbol{e}_i, \boldsymbol{e}_n|\boldsymbol{\theta}))^T,$$

para $i = 1, \ldots, n$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros do *kernel* utilizado. Assim $\boldsymbol{\theta} = [b, m]^T$ para o kernel polinomial (5.3) e $\theta = \sigma$ para o kernel gaussiano (5.4). O vetor $\boldsymbol{K}$ pode ser visto como uma linha da matriz Gram correspondente obtida com os kernels de todas possíveis combinações dos vetores de entrada estendida para incluir a possibilidade do coeficiente β_0.

Entretanto, modelando dessa forma estaríamos assumindo não existir correlação entre os elementos do vetor resposta $\boldsymbol{v}$ e que cada elemento do vetor de fluxos é independente dos outros, o que não é correto no caso do metabolismo. Na modelagem de fluxos metabólicos o problema é de resposta multivariável e as variáveis apresentam grande dependência umas das outras. Para lidar com esta situação, em (CHAKRABORTY; GHOSH; MALLICK, 2012) introduz-se um conjunto de n variáveis latentes $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ com $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^T$ de dimensão q e assume-se que os n vetores de resposta $\boldsymbol{v}_i$ são condicionalmente independentes dados $\boldsymbol{z}_i$. Além disso, assume-se que, também condicionadas em $\boldsymbol{z}_i$, as componentes de fluxo $\boldsymbol{v}_i$ são independentes entre si.

O uso de variáveis latentes nos modelos permite utilizar distribuições complexas para os dados observados representando-as como combinações de distribuições mais simples, como por exemplo, distribuições gaussianas (BISHOP, 2007). Nesse modelo, o uso de um conjunto de variáveis latentes permite modelagem tanto da correlação entre os elementos do vetor de fluxos quanto da heteroscedasticidade para estes elementos (cf. Eqs. (B.11) e (B.12)).

Com esse novo conjunto de variáveis latentes, agora é possível escrever o modelo de dados para o problema como:

$$\boldsymbol{v}_i = \boldsymbol{z}_i + \boldsymbol{\eta}_i \tag{5.6}$$

$$\boldsymbol{z}_i = \boldsymbol{K}_i^0 \boldsymbol{\beta} + \boldsymbol{\delta}_i \tag{5.7}$$

em que o vetor $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})^T$ é uma variável aleatória multivari-

ada normal de média zero e matriz de covariância $\boldsymbol{\Sigma}_\eta = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$. A matriz $\boldsymbol{K}_i^0$ é uma matriz bloco-diagonal com cada bloco igual ao vetor $\boldsymbol{K}_i$ definido anteriormente. Matematicamente podemos expressar essa matriz como o produto de Kronecker $\boldsymbol{K}_i^0 = I_q \otimes \boldsymbol{K}_i^T$, em que I_q é a matriz identidade com dimensão q. Assumindo um vetor coluna de coeficientes de regressão $\boldsymbol{\beta}_j$ para cada um dos $j = 1, \ldots, q$ elementos do vetor $\boldsymbol{v}$, o vetor de coeficientes de regressão é definido como $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_q^T]^T$. O vetor de efeitos aleatórios residuais $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iq})$ também segue uma distribuição normal de média zero e matriz de covariância $\boldsymbol{\Sigma}_\delta$, i.e. $\boldsymbol{\delta}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_\delta)$, em que $\boldsymbol{\Sigma}_\delta$ não é diagonal, de forma a correlacionar os diferentes elementos dos vetores $\boldsymbol{z}_i$. As variáveis latentes $\boldsymbol{z}_i$ e as funções $f(\boldsymbol{e}_i)$ estão relacionadas pela Equação (5.7) através do vetor $\boldsymbol{\delta}_i$. Este vetor, através da sua matriz de covariância $\boldsymbol{\Sigma}_\delta$, introduz dependência nas componentes de $\boldsymbol{z}$, o que por sua vez introduz correlação entre os elementos $\boldsymbol{v}_i$.

Os resultados na continuação do capítulo foram realizados apenas com o kernel gaussiano, pois na estimação com dados sintéticos esse modelo apresentou menor erro quadrático médio de predição.

5.2.1 Distribuições *a priori*

Para o modelo RVM da Eq.(5.6), é necessário definir distribuições *a priori* para cada um dos parâmetros e hiperparâmetros do modelo e efeitos aleatórios. A distribuição *a priori* representa a incerteza inicial (antes da observação dos dados) existente em relação a um determinado parâmetro θ do modelo (GELMAN et al., 2003). Em muitos casos, devido à necessidade de tratabilidade matemática, essas distribuições são selecionadas a partir do critério da conjugabilidade, permitindo que a distribuição *a posteriori* tenha uma solução analítica em distribuição padrão conhecida. Em outras situações, devido à falta de conhecimento sobre o processo aleatório que está sendo modelado, é possível escolher distribuições não-informativas, i.e. distribuições geralmente planas e com alta variância, que permitem atribuir valores de probabilidade semelhantes para uma grande faixa de valores do parâmetro.

Entretanto, o uso de priors não-informativos deve ser cauteloso, principalmente em modelos em que o conjunto de dados apresenta poucas observações de alta dimensionalidade.

A seleção das distribuições *a priori* para o modelo está descrita no Apêndice B, porém alguns pontos relevantes são discutidos a seguir.

Em relação aos parâmetros do *kernel* gaussiano (Eq. 5.4), a escolha da distribuição deve ser cuidadosa para não incluir a faixa de valores próximos de zero, pois valores pequenos positivos para a largura de banda no denominador (σ^2), resultam em valores do kernel também próximos de zero. Observou-se nas simulações que, caso o suporte do prior para este parâmetro inclua valores muito próximos de zero, o procedimento de estimação apresenta resultados de alta variância e muitas vezes espúrios[3].

Em relação ao kernel polinomial, observou-se que o prior discreto sugerido em (CHAKRABORTY; GHOSH; MALLICK, 2012), e incluído no Apêndice B, pode ser utilizado com segurança, porém verificamos que a utilização desse *kernel* pode comprometer a convergência das cadeias MCMC caso o valor inicial para o grau do kernel (ver Eq (5.3)) seja alto demais.

Observou-se que no caso do kernel gaussiano, o modelo RVM apresenta baixa sensibilidade a variações no parâmetro largura de banda. Priors uniformes em diferentes faixas do parâmetro não causaram alteração apreciável nos resultados do modelo. Assim optamos pelo emprego do *kernel* gaussiano.

Da mesma forma como em (CHAKRABORTY; GHOSH; MALLICK, 2012), a distribuição *a priori* para cada um dos coeficientes de regressão β_{ij} é uma distribuição normal com média zero e precisão (recíproco da variância) dada pelo respectivo hiperparâmetro λ_{ij}. Para uma determinada variável de saída, a distribuição dos coeficientes é dada por:

$$\boldsymbol{\beta}_j|\boldsymbol{\Lambda}_j \overset{\text{i.n.d}}{\sim} \mathcal{N}_{n+1}(0, \boldsymbol{\Lambda}_j^{-1}) \tag{5.8}$$

em que $j = 1, \ldots, q$ e $\boldsymbol{\Lambda}$ é a matriz de precisão. Todos os q conjuntos de coeficientes de regressão podem ser agrupados em uma matriz $\boldsymbol{\beta}$ composta de q vetores normais de tamanho $n + 1$ cada um como na Eq. (5.7). Visto que assume-se β_{ij} independentes, os hiper parâmetros $\boldsymbol{\Lambda}$ são dados por uma matriz diagonal de dimensão $(n+1)\times(n+1)$ para cada um dos $j = 1, \ldots, q$. Da mesma forma como no caso do parâmetro θ do kernel, a distribuição *a priori* para cada um dos elementos λ_{ij} é selecionada para ser não-informativa.

É interessante observar que o mecanismo de vetor de relevância automaticamente "poda"os coeficientes β_{ij} menos relevantes para a modelagem, ajustando a magnitude dos respectivos λ_{ij} para valores altos

[3] As observações iniciais dessa seção foram observadas com a utilização de *fake data*, em que define-se previamente um conjunto de parâmetros para o modelo, gerando-se os dados de saída $\boldsymbol{v}$ a partir do próprio modelo e de um conjunto de dados de entrada $\boldsymbol{e}$, realizando-se com este conjunto $\mathcal{D}$ o procedimento de estimação.

(i.e. maior precisão), concentrando a distribuição *a posteriori* do respectivo coeficiente de regressão em torno da média zero, efetivamente eliminando a participação desse coeficiente do modelo.

Uma simulação com dados sintéticos falsos (*fake data*) mostra como se dá esse comportamento dos coeficientes β_{ij} e hiperparâmetros de precisão λ_{ij}. A partir de um conjunto de $n = 20$ amostras de dados de entrada $\boldsymbol{e}$ com dimensão $p = 2000$ gerados sinteticamente, gaussianos, brancos e de média zero, foi calculado o conjunto $\boldsymbol{v}$ de dimensão $q = 12$, com um conjunto de parâmetros do modelo conhecido e pré-definido. Para essa simulação utilizou-se o kernel gaussiano com largura de banda $\theta = 13$ e coeficientes de regressão β_{ij} inteiros amostrados no intervalo $[-10, 10]$, com 20% desses iguais a zero. Os resultados são apresentados na Fig. 35. Nota-se claramente uma diferença de comportamento entre os λ_{ij} de coeficientes não-nulos e de coeficientes nulos. Os hiperparâmetros estimados para os coeficientes β_{ij} nulos (C e D) apresentam histogramas com médias altas e assimétricos. No caso dos coeficientes de regressão não-nulos (A e B), as distribuições *a posteriori* dos hiperparâmetros são inclinadas para a esquerda e com magnitude média próxima de zero, garantindo maior liberdade na estimação do correspondente parâmetro $\boldsymbol{\beta}_{ij}$.

5.3 TREINAMENTO E VALIDAÇÃO DO MODELO

5.3.1 Implementação

Inicialmente o procedimento de simulação foi implementado na linguagem de modelagem **OpenBUGS** (SPIEGELHALTER et al., 2015) *(Open* **B***ayesian Inference* **U***sing* **G***ibbs* **S***ampling)*, uma versão *open-source* multi-plataforma que atualmente substitui o conhecido pacote *WinBUGS*. Apesar da facilidade da programação dos modelos e da vantagem de utilizar distribuições pré-programadas e o total interfaceamento com R, observou-se que a complexidade do modelo sendo simulado foi excessiva para os algoritmos atualmente disponíveis no pacote *OpenBUGS*. Tempos de simulação excessivamente longos devido ao uso de amostradores genéricos e funções não-vetorizadas, além de excessiva sensibilidade a alterações em alguns parâmetros do modelo e mesmo falta de convergência exigiram a busca por uma nova solução. Durante o desenvolvimento da tese outros dois pacotes foram desenvolvidos para o mesmo tipo de aplicação, porém com vantagens em relação ao pacote *Open/WinBUGS*: **JAGS** (PLUMMER, 2015) *(Just Another*

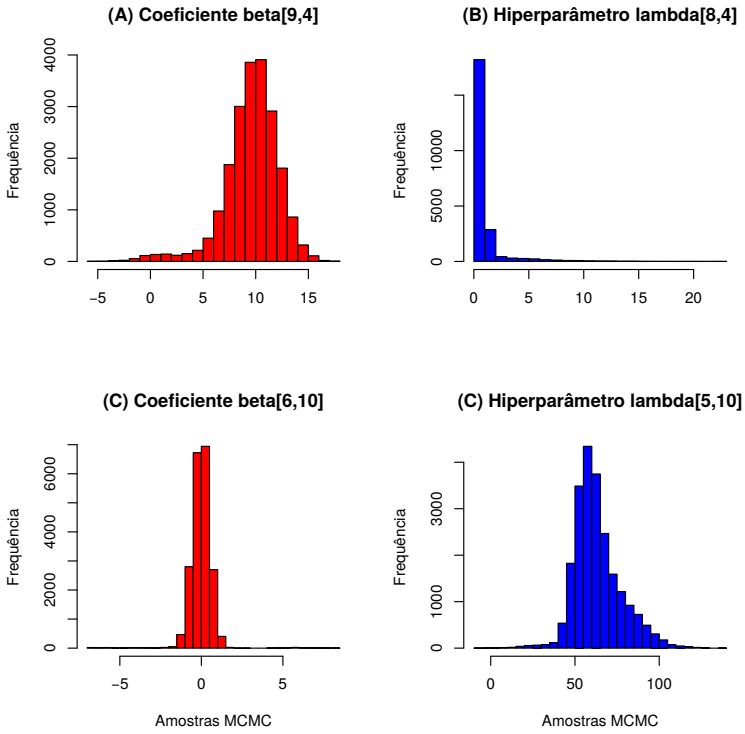

Figura 35 – Resultados das simulações MCMC para uma simulação com *fake data*. Os histogramas das amostras *a posteriori* obtidas para os coeficientes β_{ij} (vermelho) e os correspondentes histogramas dos hiperparâmetros λ (azul). Coeficiente $\beta_{9,4} = 10$, e coeficiente $\beta_{6,10} = 0$. Para o coeficiente não-nulo as amostras de $\lambda_{9,4}$ são baixas, indicando um baixo valor de precisão, permitindo o curso do valor do coeficiente correspondente até atingir o seu valor final. Para o coeficiente nulo $\beta_{6,10}$ o mecanismo automático de relevância elimina o coeficiente, aumentando a precisão ao redor da média zero. Simulação de uma cadeia MCMC com *burn-in* de 2000 amostras e total de amostragens igual a 22.000. *Fonte: Elaborada pelo autor.*

Gibbs Sampler) e **Stan** (SPIEGELHALTER et al., 2016). JAGS mantém essencialmente o mesmo paradigma de linguagem de modelagem que a usada em *OpenBUGS*, além de completo interfaceamento com R, e portanto foi a linguagem de escolha para a implementação. As maiores alterações observadas foram em relação aos tempos de simulação, que foram reduzidos em até 50% para simulações longas. Modificações como funções vetorizadas certamente foram chave para esta melhoria. A escolha de Stan para o substituto de OpenBUGS foi tentadora, porém o paradigma de programação dos modelos é relativamente mais complexo que em JAGS, indicando no horizonte uma curva de aprendizado relativamente íngreme e, portanto, não aconselhável para um trabalho em andamento. Todavia, para trabalhos de inferência bayesiana com amostradores de Gibbs que estão em seu início, essa parece ser a linguagem mais adequada e portanto recomendada. A interface de Stan com R, embora relativamente atrasada em relação a JAGS já está disponível (MCELREATH, 2016). Todos os diversos scripts de controle de simulação foram realizados na linguagem de programação R.

5.3.2 Conjunto de dados

O modelo foi inicialmente treinado e validado com um conjunto de dados sintéticos, tanto para expressão genética quanto para fluxos metabólicos. Para os dados de entrada $\boldsymbol{e}$, utilizando a estimativa de *shrinkage* para a matriz de covariância apresentada no Capítulo 2 (estimada a partir de um conjunto de dados de expressão genética do *Mycobacterium tuberculosis* com exposição à mefloquina), foram geradas $n = 100$ amostras multivariadas de vetores com a mesma dimensão dos dados originais, i.e. $p = 3979$ valores de expressão. A título de ilustração, a Figura 36 mostra, para o gene *Rv0097*, o logaritmo base 2 de alguns valores de expressão genética real das micromatrizes (utilizados na estimação da matriz de covariância) e dados sintéticos gerados.

Os dados de fluxo metabólico $\boldsymbol{v}$ correspondentes foram gerados com o método FBA apresentado no Capítulo 3, embora outro método, como E-flux (COLIJN et al., 2009), também pudesse ser utilizado. Após finalizadas as simulações de FBA, o conjunto de dados $\mathcal{D} = \{\boldsymbol{e}_i, \boldsymbol{v}_i\}$ para $i = 1 \ldots n$ foi utilizado para treinamento e validação do modelo de acordo com as indicações da próxima seção.

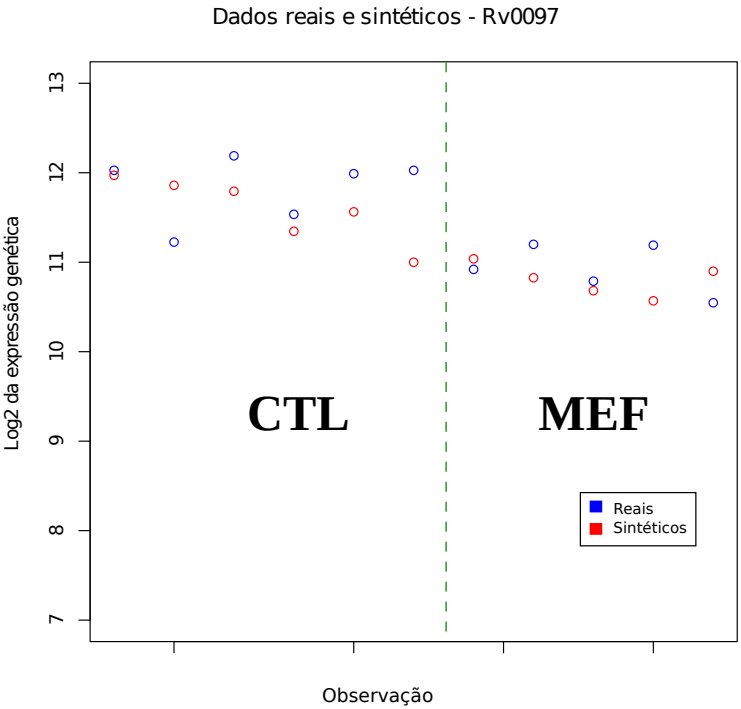

Figura 36 – Dados sintéticos e dados reais de expressão genética para o gene *Rv0097*. *Fonte: Elaborada pelo autor.*

5.3.3 Simulações

Para treinamento e validação do modelo, visto que a distribuição *a posterior* dos parâmetros do modelos é complexa demais para uma solução analítica, utilizou-se técnicas de amostragem MCMC (ROBERT; CASELLA, 2004).

Em todas as simulações MCMC foram amostradas três cadeias de Markov com inicializações independentes para garantir que o espaço da distribuição desejada seja amostrado de forma extensiva. As cadeias são utlizadas em sua totalidade, sem aplicação de *thinning* (LINK; EATON, 2012). Para determinação do período de *burn-in*, para descarte das amostras no período transitório de convergência, simulações de teste foram executadas e gráficos de convergência foram monitorados utilizando as funções do pacote R `mcmc`. Períodos de *burn-in* foram estabelecidos em $M_b = 25.000$, e as simulações foram realizadas para um total de $M = 125.000$ amostras.

Todas as simulações foram realizadas tanto com o kernel gaussiano quanto com o kernel polinomial. Os parâmetros das distribuições *a priori* foram selecionados seguindo as diretrizes indicadas em (CHAKRABORTY; GHOSH; MALLICK, 2012) para distribuições *a priori* não-informativas:

- $\gamma_1 = 2.0$, $\gamma_2 = 2.0$ (cf. Eq. (B.22))
- $c = 10^{-5}$, $d = 10^{-6}$ (cf. Eq. (B.20))
- $\psi = q + 1$, $\boldsymbol{Q} = I_q$ (cf. Eq. (B.21))

em que I_q é a matriz identidade com dimensão $q \times q$, igual ao número de elementos no vetor de saída. Para o parâmetro do kernel gaussiano utilizou-se um prior uniforme truncado de 50 e 500, com o intuito de evitar valores próximos de zero para a largura de banda. Diversas simulações iniciais tentativas foram realizadas para identificar a faixa de valores mais adequada para este prior. É importante notar que o modelo é relativamente pouco sensível a variações no valor desse parâmetro, e um estudo mais detalhado pode ser feito sobre a necessidade de estimação desse parâmetro, ao invés de um ajuste fixo com estimativa obtida por validação cruzada. No caso do kernel polinomial definiu-se um prior uniforme discreto com variável aleatória discreta na faixa $[1 \dots C]$, em que $C = 10$, sendo que cada valor pode ser escolhido com probabilidade igual a $1/C$.

Visto que o tempo de simulação torna-se excessivo caso sejam utilizados os fluxos de todas as 849 reações do modelo GSMN-TB. Por

isso, foram selecionadas as reações de 23 caminhos metabólicos que acreditamos ser mais significativos para o estudo do metabolismo do *Mycobacterium tuberculosis*. Estes caminhos foram selecionados a partir de uma análise de genes com maior expressão genética diferencial utilizando o pacote R `limma` e indicativos de caminhos afetados pela mefloquina sugeridos pelos escores Pathmod do Capítulo 4. Com um total de 207 reações, dentre esses estão incluídos caminhos metabólicos de biossíntese de lipídeos e ácidos graxos, biossíntese de biotina e MAP[4]. Ainda metabolismo de pirimidinas e purinas, degradação de benzoato e compostos como geraniol e naftaleno, respiração celular, e síntese de cofatores como vitamina B6 e cobalamina.

Para avaliar o desempenho do modelo com os dois tipos de *kernel*, treinamento e validação foram realizados 20 vezes com diferentes grupos de dados. Do conjunto de dados sintéticos com $n = 100$ amostras foram tomadas aleatoriamente $N_t = 20$ amostras para treinamento e $N_v = 20$ amostras para validação/teste. Para esse conjunto de dados realizou-se o treinamento do modelo e a predição dos fluxos metabólicos $\boldsymbol{v}$. Calculou-se o erro quadrático médio de predição (MSEP - *mean square error of prediction*) para os vetores de fluxo obtidos no procedimento de validação (comparando com os valores do conjunto de dados). O MSEP é calculado como a norma da diferença entre o vetor do conjunto de teste $\boldsymbol{v}'_i$ e o vetor $\hat{\boldsymbol{v}}_i$ resultado da predição do modelo:

$$\xi_i = ||\hat{\boldsymbol{v}}_i - \boldsymbol{v}'_i|| \quad \text{para} \quad i = 1, \ldots, N_v \tag{5.9}$$

O valor médio do erro quadrático é obtido como:

$$\bar{\xi} = \frac{1}{N_v} \sum_{i=1}^{N_v} \xi_i^2 \tag{5.10}$$

O procedimento de treinamento e validação foi realizado 20 vezes para diferentes subconjuntos de $\mathcal{D}$. A distribuição do MSEP para estes dois modelos (kernel gaussiano e kernel polinomial) estão apresentados na Figura 37. Observa-se que, embora não de forma consistente, o uso do kernel gaussiano apresenta uma melhor qualidade na predição de fluxos desconhecidos do conjunto de validação para esse problema de estimação do metabolismo a partir de dados de transcriptômica. É interessante notar que esse resultado é contrário ao apresentado em (CHAKRABORTY; GHOSH; MALLICK, 2012). Esse resultado é um indicativo de que a escolha do tipo de kernel é dependente do tipo e dimensão

[4] *Mycolic Acid Pathway* - metabolismo de ácidos micólicos

do problema (i.e. tipo dos dados) que é modelado.

Na próxima seção aplicamos o modelo a um conjunto de dados reais de experimentos de exposição do *Mycobacterium tuberculosis* à mefloquina.

5.4 APLICAÇÃO DO MODELO

Um modelo estatístico está na base de qualquer inferência ou decisão. No desenvolvimento de um modelo estatístico é necessário (1) especificar o modelo matemático, (2) ajustar o modelo aos dados e (3) interpretar os resultados. O objetivo do modelo é ser uma representação de uma população de indivíduos, para que seja possível, a partir do modelo, identificar características de interesse sobre os indivíduos dessa população. Dessa forma, os modelos estatísticos não são afirmações sobre as amostras, mas sobre a população que gerou a amostra (SINGER; WILLETT, 2003). Para que esse objetivo seja satisfeito, o processo de especificação e ajuste de um modelo deve utilizar a informação contidas nos dados experimentais, os quais são limitados, de forma a obter inferências que estejam o mais próximo possível do comportamento real da população. Nesta seção treinamos o modelo já especificado e para aplicá-lo a dados experimentais reais do *Mycobacterium tuberculosis* exposto a um candidato a composto anti-tubercular.

Para a simulação com dados reais, treinou-se o modelo com um conjunto de $n = 23$ vetores $\boldsymbol{e}$ de dados de expressão genética obtidos com experimentos de micromatrizes de DNA separados em duas condições experimentais: 12 vetores de tratamento com exposição do *Mycobacterium tuberculosis* a uma dose de $4\times$ MIC (ver (DANELISHVILI et al., 2005)) mefloquina e 11 vetores de controle. Um dos experimentos de controle apresenta alto ruído de medição segundo as métricas de qualidade Affymetrix (GAUTIER et al., 2004) e foi por isso descartado do conjunto para evitar comportamento de *outlier*. Mais uma vez, por não estarem disponíveis medições experimentais de fluxo metabólico associadas com cada um dos experimentos de micromatriz, os dados $\boldsymbol{v}$ para treinamento do modelo foram calculados com FBA. Condições de controle foram estimadas com a função objetivo de biomassa e condições de mefloquina com o método apresentado em (MONTEZANO et al., 2015). O treinamento foi realizado com o método *leave-one-out*, obtendo a predição de fluxos metabólicos $\boldsymbol{v}$.

Para cada condição experimental foram geradas pela distribuição preditiva *a posteriori* um total de $N_v = 150$ amostras. Realizou-se

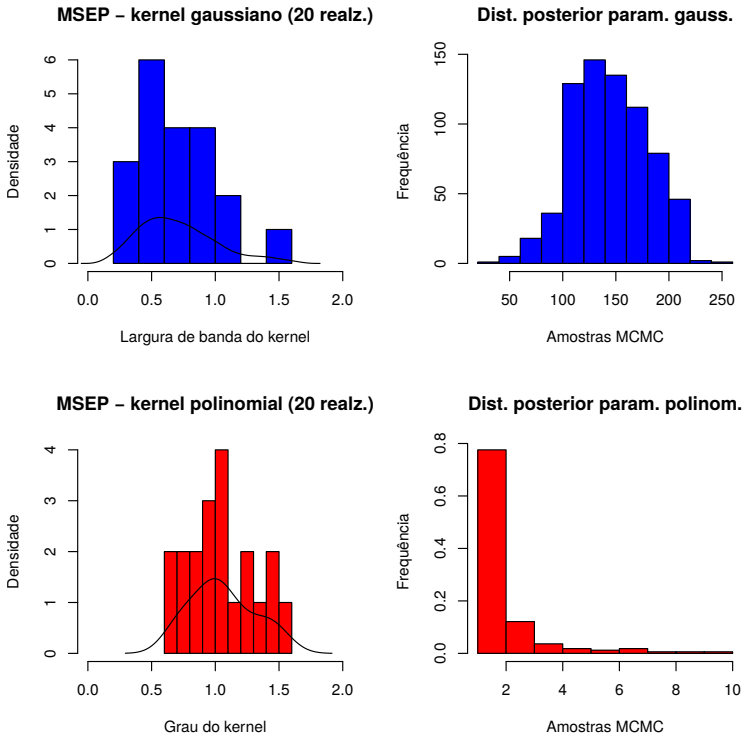

Figura 37 – Erro quadrático médio de predição para o modelo de predição do metabolismo do *Mycobacterium tuberculosis* a partir de medidas de transcriptômica. Modelo com kernel gaussiano (esq.) apresenta distribuição com menor MSEP para 20 realizações do procedimento de estimação do que para o modelo com kernel polinomial. Para este problema observa-se que o kernel gaussiano é uma melhor opção que o kernel polinomial, garantindo em geral menor MSEP. *Fonte: Elaborada pelo autor.*

um teste Welch para diferença das médias entre os fluxos metabólicos da condição de controle e da condição de mefloquina associada. A escolha do teste Welch é necessária pois não há garantia da igualdade das variâncias. Para o caso de variâncias iguais, o teste Welch produz resultados suficientemente próximos do teste-t para amostras que não sejam pequenas demais.

Os resultados do teste para cada um dos fluxos metabólicos simulados foram classificados de acordo com o valor-p resultante, que indica fluxos com a maior variação (i.e. separação das médias amostrais) entre a condição de controle e a condição de tratamento. Note-se que os fluxos podem também ser ordenados pela magnitude do fluxo médio, caso o interesse seja em encontrar fluxos com maior variação de maior magnitude. Nos resultados apresentados a seguir, apenas a variação entre as médias (independentemente da magnitude do fluxo) foi usada como métrica de seleção.

Selecionando os três caminhos metabólicos mais representativos, de acordo com as reações metabólicas apresentando as maiores variações de uma condição de controle para tratamento, observa-se que estas são reações que participam dos caminhos metabólicos *mtu00550* (síntese de peptidoglicanos), *mtu00780* (Biossíntese de biotina) e *mtu00473* (metabolismo de D-alanina).

É interessante notar que esses três caminhos são também identificados na assinatura metabólica da mefloquina obtida no Capítulo 4 com os escores de atividade Pathmod.

Embora esse resultado não identifique o mecanismo de ação do antibiótico sobre o *Mycobacterium tuberculosis*, ainda assim permite que obtenha-se uma direção de pesquisa para busca de novos alvos metabólicos para compostos antibióticos.

5.5 CONCLUSÕES

Neste capítulo apresentamos um modelo estatístico para modelagem do metabolismo na forma de um conjunto de fluxos metabólicos. As configurações de fluxo são estimadas a partir de dados de transcriptômica.

O modelo de máquina de vetor de relevância estudado é adequado para o problema de modelagem do metabolismo pois reduz a dimensionalidade do problema e ao mesmo tempo faz uso de um número pequeno de amostras para reduzir o número de parâmetros a estimar. O uso de *kernels* permite modelar relações entrada-saída com alto grau

de não-linearidade enquanto atua na redução da dimensão do problema. O modelo é multivariado e permite a estimação de correlações entre as variáveis de saída (i.e. fluxos metabólicos, além de modelar a heteroscedasticidade do vetor de fluxos (fluxos com variâncias diferentes). O uso de um modelo com formulação bayesiana fornece uma maneira acessivel para incorporação de conhecimento biológico no problema ou modelagem de incerteza sobre os parâmetros do modelo.

O modelo RVM permite o uso de variados *kernels*, e neste trabalho realizou-se uma simulação com dados sintéticos que permitiu verificar que, para o problema da modelagem do metabolismo, o kernel gaussiano apresenta menor erro de predição que o kernel polinomial e, portanto, apresenta melhor capacidade de modelagem da relação entre dados de transcriptômica e fluxos metabólicos.

O uso do modelo com dados experimentais reais de expressão genética do *Mycobacterium tuberculosis* exposto à mefloquina permitiu obter um conjunto de amostras da distribuição preditiva *a posteriori* dos fluxos de diversos caminhos metabólicos para as duas condições experimentais: controle e tratamento com mefloquina. Com um teste Welch de diferença de médias para variâncias desiguais foi possível classificar os fluxos da predição de acordo com o valor-p do teste. Os testes com menor valor-p apresentam as diferenças mais significativas entre as médias do CTL e TRT.

Na lista das reações identificadas estão presentes diversas reações de caminhos metabólicos associados com a síntese de componentes da membrana e parede celular, mostrando sobreposição com os resultados encontrados para a mefloquina com os escores Pathmod e apresentados no Capítulo 4. Os caminhos metabólicos com maior quantidade de reações com maior variação de fluxo metabólico são os caminhos de biossíntese de peptidoglicano, síntese de biotina e metabolismo de d-alanina, indicando que a maior atividade diferencial metabólica ocorre em processos da membrana e parede celular, como por exemplo a produção dos ácidos micólicos.

Os resultados deste capítulo mostram que é possível a modelagem do metabolismo a partir de dados de transcriptômica com um modelo estatístico com capacidade de modelagem não-linear e que possa ser estimado mesmo em situações de dados com alta dimensionalidade e poucas amostras (problemas n-pequeno, p-grande).

Em termos de implementação, embora o uso da linguagem de modelagem JAGS permita simulações significativamente mais rápidas que outros pacotes de inferência bayesiana como OpenBUGS, ainda assim os tempos de simulação se tornam excessivamente longos à me-

dida que deseja-se estimar mais fluxos metabólicos (maior dimensão q). Para as simulações deste capítulo, ainda que os fluxos de apenas duas dezenas de caminhos metabólicos tenham sido contempladas na simulação, ainda os tempo de simulação são longos, com as simulações de dados reais e validação do tipo *leave-one-out*, necessitando de vários dias para serem completadas.

Em geral, a maior dificuldade dos modelos complexos utilizando técnicas de amostragem MCMC são os tempos de simulação longos, devido à necessidade de longos períodos para convergência das cadeias amostradoras. O problema á agravado no caso de modelos multivariados, em que a quantidade de processamento aumenta, não necessariamente linearmente, com o aumento do número de variáveis a serem simuladas. Entretanto, no caso da RVM, esse problema está associado apenas ao número q de variáveis de saída, visto que, no caso dos dados de entrada $\boldsymbol{e}$, esses são processados pelo kernel e sua dimensionalidade não afeta de forma significativa a complexidade computacional do modelo.

6 CONCLUSÃO

6.1 CONSIDERAÇÕES FINAIS

A temática deste trabalho multidisciplinar foi desenvolver e avaliar técnicas computacionais para a modelagem de metabolismo, mais especificamente o metabolismo do organismo *Mycobacterium tuberculosis*. As técnicas e modelos apresentados trazem contribuições similares a outros métodos *in silico* disponíveis na literatura (CHANDRASEKARAN; PRICE, 2010; COLIJN et al., 2009; BESTE et al., 2007; EFRONI; SCHAEFER; BUETOW, 2007).

Um dos objetivos do trabalho foi estudar técnicas que pudessem ser úteis em aplicações como o desenvolvimento de compostos antibióticos para o *Mycobacterium tuberculosis*, todavia, não foi objetivo do trabalho identificar estes antibióticos ou o seu mecanismo de ação, nem tampouco sugerir sua aplicabilidade como possíveis compostos candidatos a drogas anti-tuberculose.

O objetivo principal do trabalho foi desenvolver, apresentar e avaliar um número de técnicas computacionais, avaliar a sua adequação para estudo do metabolismo e indicar possíveis vias de utilização destas ferramentas. Nenhuma dessas ferramentas foi implementada pensando em sua utilização comercial ou por profissionais com pouca familiaridade com técnicas computacionais. Todavia, seria de grande valia que os métodos aqui apresentados pudessem ser desenvolvidos, formatados e adaptados e disponibilizados para a comunidade científica. É crença do autor que os métodos podem se tornar ferramentas úteis no dia-a-dia do investigador envolvido em estudos do metabolismo, pesquisa e desenvolvimento de compostos antibióticos para os mais diversos organismos patogênicos, incluído aí o *Mycobacterium tuberculosis*.

As técnicas computacionais apresentadas em geral foram tomadas de modelos pré-existentes de forma a aprimorá-los (Capítulo 2), complementá-los (Capítulo 3), adaptá-los (Capítulo 4) ou avaliá-los (Capítulo 5) para o estudo do metabolismo. Os modelos de sistemas biológicos são complexos e não é possível incluir em um modelo toda a sua complexidade, a qual não é nem mesmo conhecida em toda sua complexidade. Os modelos apresentados são como pequenas lentes que permitem observar estes sistemas de forma bastante limitada e mesmo pouco nítida. Ainda assim, são úteis, pois as inferências que eles permitem guiam no dia-a-dia pesquisadores na direção correta de novas descobertas sobre os organismos. O desenvolvimento de novos modelos

deve contemplar os sucessos e falhas observados em tentativas precedentes, em um processo iterativo constante, onde dados experimentais reforçam ou refutam as predições e conclusões obtidas dos modelos *in silico*, forçando serem repensados e aprimorados, e em que as predições dos modelos sugerem novos tipos de experimentos que podem ser realizados e que ainda não haviam sido imaginados.

Embora dados experimentais tenham sido utilizados nos modelos do presente trabalho, os resultados apresentados não foram validados por novos experimentos biológicos de bancada especificamente projetados para permitir afirmar ou refutar as predições e conclusões indicadas. Todas as validações foram realizadas com dados experimentais já disponíveis na literatura e comparados com resultados obtidos com outros métodos *in silico* já publicados. Necessariamente, projetar novos experimentos que permitam validar os modelos apresentados é um próximo passo importante na continuidade deste trabalho. Mostrar que os indicativos obtidos com os escores Pathmod refletem as tendências da atividade metabólica do organismo, ou que as distribuições de fluxo obtidas com FBA e a função objetivo de proteômica realmente produzem resultados de fluxo mais próximos que com a função objetivo de biomassa no caso de exposição do organismo a antibióticos, ou ainda que a lista de caminhos metabólicos e reações metabólicas com maior atividade diferencial obtidos com o modelo estatístico realmente podem indicar vias de acesso a alvos racionais para novos compostos antibióticos ou uma melhor compreensão dos seus mecanismos de ação, esses são passos de validação importantes que devem ser realizados para comprovar a utilidade das técnicas apresentadas.

Certamente existem maneiras alternativas de complementar a validação dos métodos. Neste trabalho estudou-se especificamente o metabolismo do organismo *Mycobacterium tuberculosis*, pois originalmente essa foi a ideia motivadora de todo o trabalho. Entretanto, o *Mycobacterium tuberculosis* não é considerado um organismo modelo para a biologia (devido a sua alta virulência), e dessa forma a quantidade de informações disponíveis sobre a sua fisiologia ou metabolismo ou caminhos de sinalização celular, não pode ser comparada a, por exemplo, o organismo *Escherichia coli*. Uma proposta de validação interessante para todos os métodos apresentados pode ser o uso de conjuntos de dados (disponíveis em quantidades e variedades muito maiores que para o Mtb) e redes metabólica e regulatória deste organismo-modelo para analisar os resultados obtidos. De qualquer forma, ainda que validações (positivas ou negativas) possam ser obtidas dessa forma, a confiabilidade no em qualquer método ou modelo será significativamente

aumentada à medida que suas predições confirmarem dados observados em experimentos independentes. De qualquer forma, observou-se que ferramentas com aplicabilidade a uma variedade de organismos às vezes são vistas com ceticismo por determinadas áreas da comunidade científica. Portanto, a aplicação e validação de técnicas computacionais para um organismo necessariamente não significa que a sua validação possa ser transposta e assumida para outros organismos. O autor deste trabalho não possui conhecimentos suficientes para poder determinar as razões reais por trás dessa dificuldade, mas é importante tê-la em mente no caso de uso de uma ferramenta ou modelo computacional com um organismo para o qual a validação desta ferramenta não foi realizada.

Durante este trabalho observou-se que em muitas instâncias os modelos e técnicas computacionais são utilizadas por investigadores de forma relativamente *ad hoc* e acrítica. Por exemplo, no Capítulo 2 avaliamos o uso da técnica de encolhimento na estimação da matriz de covariância de dados de alta dimensionalidade de expressão genética. Embora nestas situações a estimativa de máxima verossimilhança não possa ser aplicada com suficiente confiabilidades devido ao baixo número de amostras disponíveis para sua estimação, os autores em (SCHÄFER; STRIMMER, 2005) observaram que em diversos artigos publicados, essa é exatamente a estimativa utilizada, retirando conclusões e inferências a partir dessas estimativas contendo essencialmente nenhuma informação sobre o processo aleatório subjacente, devido a sua alta variância e pobre condicionamento.

Por esse motivo, é importante que trabalhos multidisciplinares deste tipo, que desenvolvem técnicas computacionais para nichos de aplicação onde os usuários podem ser leigos sobre as minúcias e implicações dos algortimos, suposições assumidas, hipóteses estatísticas assumidas e mesmo adequação de um método a determinado problema, não se observe apenas o desenvolvimento da ideia do método. É necessário que se leve em consideração os usuários das técnicas. É necessário que se leve em consideração as necessidades destes usuários que poderão ser satisfeitas com determinada ferramenta computacional. O desenvolvedor de métodos computacionais deve poder indicar como e quando tal método pode e deve ser utilizado, quais suas limitações e em que medida os seus resultados podem ser interpretados e que tipo de inferências e conclusões podem ser derivadas destes resultados. Certamente estas dificuldades não estão apenas associadas a trabalhos na área da Biologia Computacional, estendendo-se a um número de diversas outras áreas onde o rigor matemático não é a primeira preocupação

do pesquisador. Todavia, se desejamos que a área da Biologia Computacional seja beneficiada pelos métodos matemáticos e algoritmos que tem sido desenvolvidos nos últimos anos, e que tais ferramentas atinjam o objetivo a que se propõem junto aos usuários finais e contribuam de forma significativa para o avanço da área, é importante ter em mente tais preocupações.

Em suma, o trabalho apresentado nesta tese mostra que é possível ampliar, reciclar e adaptar técnicas computacionais existentes para estudo do metabolismo de organismos procariotas agentes de doenças. Na seção seguinte discutimos três áreas possíveis, dentre várias, que podem ser utilizadas para continuidade deste trabalho.

6.2 PROPOSTAS DE TRABALHOS FUTUROS

Validação biológica. Este ponto já foi mencionado na seção anterior. Modelos computacionais são úteis na medida que estão harmonizados com o funcionamento real dos organismos, ainda que de forma simplificada e incompleta. Entretanto, essa sintonia só pode ser obtida quando os resultados computacionais são validados e confirmados por experimentos biológicos de bancada. Neste trabalho foi realizada a parte de desenvolvimento e avaliação computacional dos modelos, necessitando ainda de validação biológica.

Entretanto, nem sempre essas validações são simples de realizar. Especialmente no caso do metabolismo, dados experimentais de validação devem ser de algum tipo que permita observar variações ou intensidade de fluxos metabólicos, um procedimento que não é simples de realizar e com as te?nicas existentes, não consegue ser compreensivo. Ao passo que em um modelo de rede metabólica seja possível especificar as diversas reações metabólicas de uma determinada função do organismo, sua estequiometria e mesmo as enzimas que catalisam cada uma dessas reações e os genes regulatórios dessas enzimas, como foi feito na Capítulo 4, no momento que estes precisam ser validados biologicamente, seria necessário realizar para uma mesma condição experimental medidas transcriptômicas, proteômicas, de fluxoma e mesmo metaboloma para que fosse possível avaliar predições de tal modelo. Isso não é factível. No caso de medições de fluxo metabólico, em geral estas são realizadas de forma indireta contabilizando variação nos metabólitos processados, o que permite avaliação de apenas um subconjunto de fluxos intracelulares. É possível que com o avanço tecnológico das últimas décadas e a crescente aceleração desse avanço, em breve estejam dis-

poníveis métodos de análise do fluxoma mais completas, o que não é possível atualmente.

Uso de conhecimento *a priori*. No modelo apresentado no Capítulo 5 foi utilizada uma técnica bayesiana, cuja grande vantagem é a possibilidade de incorporação de conhecimento biológico existente na forma de distribuições *a priori*. Esse estudo não foi realizado, porém seria altamente recomendável em uma possível continuidade deste trabalho. A inclusão de conhecimento *a priori* permitirá diminuir a variância das estimativas do modelo, compensando de forma mais incisiva o número reduzido de amostras disponíveis para a estimação do modelo. Convergência das cadeias amostradoras MCMC certamente seria beneficiada com a inclusão desse conhecimento.

Todavia, essa questão necessariamente passa por uma discussão com profissionais biólogos que possuem experiência com o organismo e podem identificar qual a incerteza e nível de conhecimento sobre cada aspecto do modelo onde esse conhecimento poderia ser incluído. No caso do modelo RVM apresentado existe possibilidade de modificar os priors associados com os coeficientes da regressão, os efeitos aleatórios dos fluxos de saída, incluindo sua heteroscedasticidade intrínseca, e correlações entre os diversos fluxos meabólicos. Os priors para os coeficientes de regressão em geral não podem ser modificados de forma útil, visto que sua interpretação biológica é dificultada pelo uso de kernels, que transforma não-linearmente os dados de transcriptômica de entrada.

Um estudo sobre essa possibilidade, ou mesmo uma análise sobre possíveis priors não-informativos mais adequados que os utilizados neste trabalho seria uma desejável linha de pesquisa para a continuação. Visto que o modelo é estimado com técnicas de amostragem MCMC, não há real necessidade que os priors tenham propriedade de conjugação, podendo ser sacrificada a tratabilidade matemática. Dessa forma, poderia ser colocada em questão o uso de priors gaussianos para os coeficientes de regressão, priors uniformes para os parâmetros do kernel, priors gama-inversa para os hiperparâmetros do mecanismo de relevância, priors gama para as variâncias dos fluxos, etc.

Implementação e desenvolvimento de pacotes de software. Todos os métodos apresentados neste trabalho foram implementados na linguagem de programação R. Essa escolha se deu por características da linguagem, que embora atualmente, de propósito geral, foi desenvolvida especificamente para estudos estatísticos, além de ótimas características de *scripting* que facilitam o seu interfaceamento com diversos

outros pacotes, linguagens e ferramentas do ambiente Linux.

Através do portal Bioconductor, é possível disponibilizar estes métodos para a comunidade científica sem custo, produzindo visibilidade para a pesquisa e auxiliando no desenvolvimento de ferramentas de qualidade. Necessriamente, os scripts destes métodos precisam ser aprimorados e reprogramados com a intenção de melhorar o seu desempenho e seu uso em ambientes computacionais diversos. Uma vez realizada essa adaptação (que pode ser feita de forma iterativa ao longo do processo de desenvolvimento do pacote), imediatamente os métodos estariam disponíveis para a comunidade científica mundial, recebendo realimentação sobre os resultados obtidos, as diferentes aplicações em que a ferramenta será utilzada, e o mais importante, informações de como o software pode ser melhorado e aprimorado. Esta proposta de continuidade do trabalho é extremamente atraente, e deveria ser considerada para um trabalho futuro.

Em relação ao método do Capítulo 5, uma possibilidade de melhoria altamente recomendada seria a substituição da linguagem de modelagem JAGS pela linguagem de programação Stan, mencionados brevemente na seção Implementação deste capítulo. Stan surge como uma linguagem bastante poderosa para inferências em modelos bayesianos, com foco principalmente em modelos complexos, e seu aprendizado parece estar na ordem do dia para qualquer profissional que queira trabalhar na área.

Pelas experiências observadas neste trabalho, a linguagem OpenBUGS deveria ser desconsiderada como alternativa, a não ser para estimação e inferências em modelos de complexidade relativamente baixa. A linguagem JAGS demonstrou ser uma alternativa significativamente mais poderosa em relação a OpenBUGS, com desempenho muito superior, principalmente em relação aos tempos de simulação muito longos. Seu uso em modelos iniciais é altamente recomendável pela simplicidade da linguagem de modelagem e curva de aprendizado relativamente pouco inclinada e curta. Entretanto, é possível que o seu desempenho para modelos computacionais muito complexos possa ser superado com os algoritmos de amostragem mais eficientes disponíveis em Stan.

REFERÊNCIAS

ALBERTS, B. et al. **Molecular Biology of the Cell**. New York, NY: Garland Science, 2002.

ALMHANA, J. et al. A recursive algorithm for gamma mixture models. In: **2006 IEEE International Conference on Communications**. Istambul, Turquia: [s.n.], 2006. v. 1, p. 197–202.

ALTSCHUL, S. et al. Basic local alignment search tool. **J Mol Biol**, v. 215, n. 3, p. 403–10, Out 1990.

APWEILER, R. et al. UniProt: the universal protein knowledgebase. **Nucl. Ac. Res.**, v. 32, n. 1, p. D115–D119, Fev 2004.

BALDI, P.; LONG, A. D. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. **Bioinformatics**, v. 17, n. 6, p. 509–19, Jun 2001.

BALÁZSI, G. et al. The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. **Molecular Systems Biology**, v. 4, n. 225, p. 1–8, Set 2008.

BAR-JOSEPH, Z. Analyzing time series gene expression data. **Bioinformatics**, v. 20, n. 16, p. 2493–503, Mai 2004.

BESTE, D. et al. GSMN-TB: a web-based genome scale network model of *Mycobacterium tuberculosis* metabolism. **Genome Biology**, BioMed Central, v. 8, n. 5, p. R89.1–R89.17, 2007.

BISHOP, C. **Pattern Recognition and Machine Learning**. New York, NY: Springer, 2007.

BOSHOFF, H. I. et al. The transcriptional responses of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action. **J Biol Chem**, v. 279, n. 38, p. 40174–84, Set 2004.

CAMPBELL, N. A.; REECE, J. B. **Campbell Biology**. 9. ed. San Francisco, CA: Pearson Education, 2009.

CASTELNUOVO, B. A review of compliance to anti tuberculosis treatment and risk factors for defaulting treatment in Sub Saharan Africa. **African Health Sciences**, African Health Journals Partnership Project, v. 10, n. 4, p. 320–4, Dez 2010.

CERDEIRA, J. O. et al. **subselect: Selecting Variable Subsets**. [S.l.], 2015. R package version 0.12-5. Disponível em: <http://CRAN.R-project.org/package=subselect>.

CHAKRABORTY, S.; GHOSH, M.; MALLICK, B. K. Bayesian nonlinear regression for large p small n problems. **J. Multivariate Analysis**, v. 108, n. 1, p. 28–40, Jul 2012.

CHANDRASEKARAN, S.; PRICE, N. D. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. **Proceedings of the National Academy of Sciences of the United States of America**, PNAS, v. 107, n. 41, p. 17845–50, Set 2010.

COLIJN, C. et al. Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. **PLoS Comput. Biol**, v. 5, n. 8, p. e1000489, Ago 2009.

CONGDON, P. **Bayesian Statistical Modelling**. 2. ed. Sussex, UK: John Wiley & Sons, 2006.

COVERT, M. et al. Integrating high-throughput and computational data elucidates bacterial networks. **Nature**, v. 429, n. 6987, p. 92–6, Mai 2004.

COVERT, M.; SCHILLING, C. H.; PALSSON, B. O. Regulation of gene expression in flux balance models of metabolism. **Journal of Theor. Biol.**, v. 213, n. 1, p. 73–88, Nov 2001.

DANELISHVILI, L. et al. Genomic approach to identifying the putative target of and mechanisms of resistance to mefloquine in mycobacteria. **Antimicrob. Agents Chemotherapy**, v. 49, n. 9, p. 3707–14, Set 2005.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **Nucl. Ac. Res.**, v. 30, n. 1, p. 207–10, Out 2002.

EFRONI, S. et al. The PathOlogist: an automated tool for pathway-centric analysis. **BMC Bioinformatics**, v. 12, n. 133, p. 1–10, Jan 2011.

EFRONI, S.; SCHAEFER, C.; BUETOW, K. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. **PLoS One**, v. 2, n. 5, p. e425, Mai 2007.

FLOYD, K.; PANTOJA, A.; DYE, C. Financing tuberculosis control: the role of a global financial monitoring system. **Bulletin of the World Health Organization**, v. 85, n. 5, p. 325–420, Mai 2007.

FOLEY, M.; TILLEY, L. Quinolone antimalarials: mechanisms of action and resistance. **Int. J. Parasitol.**, v. 27, n. 2, p. 231–40, Fev 1997.

FRANKEL, B. A.; BLANCHARD, J. S. Mechanistic analysis of *Mycobacterium tuberculosis* rv1347c, a lysine n^{ϵ}-acyltransferase involved in mycobactin biosynthesis. **Arch Biochem Biophys**, v. 477, n. 2, p. 259–66, Set 2008.

Fred Hutchinson Cancer Research Center. **Bioconductor - Open Source Software for Bioinformatics**. Seattle, WA, 2015. Disponível em: <http://www.bioconductor.org>.

FRICK, M.; JIMÉNEZ-LEVI, E. 2013 report on tuberculosis research funding trends, 2005—2012. **Treatment Action Group, Stop TB Partnership**, Nov 2013.

GAUTIER, L. et al. *affy* - analysis of affymetrix genechip data at the probe level. **Bioinformatics**, v. 20, n. 3, p. 307–15, Jul 2004.

GELMAN, A. et al. **Bayesian Data Analysis**. New York, NY: Chapman and Hall, 2003.

GELMAN, A.; HILL, J. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. New York, NY: Cambridge University, 2007.

GEVORGYAN, A. et al. SurreyFBA: a command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks. **Bioinformatics**, v. 27, n. 3, p. 433–4, Fev 2011.

GOTTLIEB, D.; SHAW, P. D. **Antibiotics (Mechanisms of Action)**. New York, NY: Springer Verlag, 1967.

HOFFBECK, J.; LANDGREBE, D. A. Covariance matrix estimation and classification with limited training data. **IEEE Trans. Pattern**

Analysis and Machine Intelligence, v. 18, n. 7, p. 763–7, Fev 1996.

IRIGOIEN, I.; VIVES, S.; ARENAS, C. Microarray time course experiments: Finding profiles. **IEEE/ACM Trans. Comput. Biol. Bioinformatics**, v. 8, n. 2, p. 464–75, Mar 2011.

KAERN, M. et al. Stochasticity in gene expression: from theories to phenotypes. **Nature Review: Genetics**, v. 6, n. 1, p. 451–64, Jun 2005.

Kanehisa Labs. **KEGG - Kyoto Encyclopedia of Genes and Genomes**. Kyoto, Japão, 2015. Disponível em: <http://www.kegg.jp/kegg/>.

KELES, S. Mixture modeling for genome-wide localization of transcription factors. **Biometrics**, v. 63, n. 1, p. 2118–22, Set 2007.

KHALILI, A. et al. Normal-gamma mixture model for detecting differentially methylated loci in three breast cancer cell lines. **Cancer Inform.**, v. 3, p. 43–54, Jan 2007.

KIM, B. H.; GADD, G. M. **Bacterial Physiology and Metabolism**. New York, NY: Cambridge University Press, 2008.

KUNDROTAS, P. J.; VAKSER, I. A. Accuracy of protein-protein binding sites in high-throughput template-based modeling. **PLoS Comput Biol**, v. 6, n. 4, p. e1000727, Fev 2010.

LAKSHMANAN, M. et al. Software applications for flux balance analysis. **Briefings in Bioinformatics**, v. 1, p. 1–15, Nov 2012.

LEDOIT, O.; WOLF, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. **J. Empir. Finance**, v. 10, n. 5, p. 603–21, 2003.

LEE, M.-L. T. **Analysis of Microarray Gene Expression Data**. Berlim, Alemanha: Springer Verlag, 2004.

LESK, A. M. **Introduction to Genomics**. New York, NY: Oxford University, 2007.

LEWIN, A. et al. Bayesian modeling of differential gene expression. **Biometrics**, v. 62, n. 1, p. 1–9, Mar 2006.

LINK, W. A.; EATON, M. J. On thinning of chains in MCMC. **Methods in Ecology and Evolution**, v. 3, n. 1, p. 112?5, 2012.

MACHADO, D.; HERRGARD, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. **PLoS Comput. Biol.**, v. 10, n. 4, p. e1003580, Abr 2014.

MADIGAN, M. T. et al. **Brock Biology of Microorganisms**. 12. ed. San Francisco, CA: Pearson Education, 2006.

MAHADEVAN, R.; SCHILLING, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. **Metab. Eng.**, v. 5, n. 4, p. 264–76, Out 2003.

MATSUMOTO, S. S. et al. Mechanism of ascididemin-induced cytotoxicity. **Chem Res Toxicol.**, v. 16, n. 2, p. 113–22, Fev 2003.

MATTICK, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. **BioEssays**, v. 25, n. 1, p. 930–9, Jan 2003.

MAYROSE, L.; FRIEDMAN, N.; PUPKO, T. A gamma mixture model better accounts for among site rate heterogeneity. **Bioinformatics**, v. 21, n. 2, p. 151–8, Jan 2005.

MCELREATH, R. **Statistical Rethinking: A Bayesian Course with Examples in R and Stan**. New York, NY: Chapman & Hall/CRC Press, 2016.

MCNAMARA, M. et al. The surface proteome of *Mycobacterium avium* subsp hominissuis in the early stages of macrophage infection. **Infect. Immun.**, v. 80, n. 5, p. 1868–80, Mai 2012.

MCNAMARA, M. et al. Surface-exposed proteins of mycobacteria and the role of Cu-Zn superoxide dismutase in macrophage and neutrophil survival. **Proteome Sci.**, v. 11, n. 1, p. 1–7, Nov 2013.

MONTEZANO, D. et al. Flux balance analysis with objective function defined by proteomics data - metabolism of *Mycobacterium tuberculosis* exposed to mefloquine. **PLoS ONE**, v. 10, n. 7, p. e0134014, Jul 2015.

NEWMAN, M. E. J. **Networks: An Introduction**. Oxford, UK: Oxford University Press, 2010.

NEWTON, M. et al. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. **J. Comput. Biol.**, v. 8, n. 1, p. 37–52, Jul 2004.

OPGEN-RHEIN, R.; STRIMMER, K. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. **BMC Bioinformatics**, v. 8, n. 2, p. 1–8, Jan 2007.

ORTH, J. D.; THIELE, I.; PALSSON, B. O. What is flux balance analysis? **Nature Biotechnology**, Nature Publishing Group, v. 28, n. 3, p. 245–8, Mar 2010.

OSBORNE, R. First novel anti-tuberculosis drug in 40 years. **Nature Biotechnology**, Nature America, v. 31, n. 2, p. 89–91, Fev 2013.

PALSSON, B. O. The challenges of *in silico* biology. **Nature Biotechnology**, v. 18, p. 1147–50, Nov 2000.

PAPADIMITRIOU, C. H.; STEIGLITZ, K. **Combinatorial Optimization: Algorithms and Complexity**. Englewood Cliffs, NJ: Prentice Hall, 1982.

PAULSSON, J. Models of stochastic gene expression. **Physics of Life Reviews**, v. 1, p. 157–75, Mai 2005.

PAVIA, D. L. et al. **Introdução à Espectroscopia - tradução da 4ª edição norte-americana**. Brasil: Cengage Learning, 2010.

PIEDADE, R. et al. Carboxymefloquine, the major metabolite of the antimalarial drug mefloquine, induces drug-metabolizing enzyme and transporter expression activation of pregnane x receptor. **Antimicrob. Agents Chemother.**, v. 59, n. 1, p. 96–104, Out 2014.

PLUMMER, M. **JAGS - Just Another Gibbs Sampler**. [S.l.], 2015. Disponível em: <http://www.mcmc-jags.sourceforge.net>.

Proteome Software. **Proteome Software Scaffold Wikispaces — Proteomics**. Portland, OR, Dez 2014. Disponível em: <http://proteome-software.wikispaces.com/Proteomics>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Viena, Áustria, 2013. Disponível em: <http://www.R-project.org/>.

RAMAN, K.; CHANDRA, N. Flux balance analysis of biological systems: applications and challenges. **Briefings in Bioinformatics**, v. 10, n. 4, p. 435–49, 2009.

RAMAN, K.; RAJAGOPALAN, P.; CHANDRA, N. Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs. **PLoS Comput. Biol.**, v. 1, n. 5, p. e46, 2005.

REED, J. L.; PALSSON, B. O. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states. **Genome Res.**, v. 14, n. 9, p. 1797–805, Set 2004.

ROBERT, C. P.; CASELLA, G. **Monte Carlo Statistical Methods**. New York, NY: Springer, 2004.

ROESSNER, U.; BOWNE, J. What is metabolomics all about? **BioTechniques**, v. 46, n. 5, p. 363–5, Abr 2009.

RUSTAD, T. et al. Mapping and manipulating the *Mycobacterium tuberculosis* transcriptome using a transcription factor overexpression-derived regulatory network. **Genome Biology**, v. 15, n. 11, p. 502, Fev 2014.

SANZ, J. et al. The transcriptional regulatory network of *Mycobacterium tuberculosis*. **PLoS ONE**, PLoS, v. 6, n. 7, p. e22178, Jul 2011.

SANZEY, B. Modulation of gene expression by drugs affecting deoxyribonucleic acid gyrase. **J Bacteriol.**, v. 138, n. 1, p. 40–7, Apr 1979.

SASSETTI, C. M.; BOYD, D. H.; RUBIN, E. J. Genes required for mycobacterial growth defined by high-density mutagenesis. **Mol. Microbiol.**, v. 48, n. 1, p. 77–84, Abr 2003.

SCHUETZ, R.; KUEPFER, L.; SAUER, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. **Molecular Systems Biology**, Nature Publishing Group, v. 3, p. 119:1–119:15, Mai 2007.

SCHÄFER, J. et al. **corpcor: Efficient Estimation of Covariance and (Partial) Correlation**. [S.l.], 2015. R package version 1.6.8. Disponível em: <http://CRAN.R-project.org/package=corpcor>.

SCHÄFER, J.; STRIMMER, K. A shrinkage approach to large-scale covariance matrix estimation and inplications for functional genomics. **Statistical Applications in Genetics and Molecular Biology**, v. 4, n. 1, p. 32:1–30, Jan 2005.

SCHÖLKOPF, B.; SMOLA, A. J. **Learnings with Kernels**. Cambridge, MA: MIT Press, 2002.

SEGAL, E. et al. Rich probabilistic models for gene expression. **Bioinformatics**, v. 17, n. Supp. 1, p. S243–S52, Jun 2001.

SHANNON, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome Research**, v. 13, n. 1, p. 2498–504, Dez 2003.

SHLOMI, T. et al. Network-based prediction of human tissue-specific metabolism. **Nature Biotechnology**, v. 26, p. 1003–10, Set 2008.

SINGER, J. D.; WILLETT, J. B. **Applied Longitudinal Data Analysis**. Oxford, UK: Oxford University Press, 2003.

SLAYDEN, R. A.; BARRY, C. E. The genetics and biochemistry of isoniazid resistance in *Mycobacterium tuberculosis*. **Microbes and Infection**, v. 2, n. 6, p. 659–669, Jan 2000.

SPIEGELHALTER, D. et al. **OpenBUGS**. [S.l.], 2015. Disponível em: <http://www.openbugs.net>.

SPIEGELHALTER, D. et al. **Stan - Stan**. [S.l.], 2016. Disponível em: <http://www.mc-stan.org>.

STEENWINKEL, J. D. M. de et al. Time-kill kinetics of anti-tuberculosis drugs, and emergence of resistance, in relation to metabolic activity of *Mycobacterium tuberculosis*. **Journal of Antimicrobial Chemotherapy**, British Society for Antimicrobial Chemotherapy, v. 65, n. 12, p. 2582–9, Dez 2010.

Swiss Institute of Bioinformatics and École Polytechnique Fédérale de Lausanne. **Tuberculist - Mycobacterium tuberculosis Database**. Lausana, Suíça, 2013. Disponível em: <http://tuberculist.epfl.ch>.

TADJUDIN, S.; LANDGREBE, D. A. Covariance estimation with limited training samples. **IEEE Trans. Geosci. Remote Sensing**, n. 4, p. 2113–8, Out 1999.

TB Alliance. **Global Alliance for TB Drug Development**. Nova Iorque, NY, 2000. Disponível em: <http://www.tballiance.org>.

TB Database. **Find Regulators for A Set of Genes**. [S.l.], 2015. Disponível em: <http://genome.tbdb.org/tbdb_sysbio/RegulatorsFromList.html>.

TB Database. **An Integrated Platform for Tuberculosis Research**. [S.l.], 2015. Disponível em: <http://www.tbdb.org>.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. Hoboken, NJ: Wiley Interscience, 2008.

THIELE, I.; PALSSON, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. **Nature Protocols**, v. 5, n. 1, p. 93–121, Jan 2010.

THOMAS, R. et al. Validation and characterization of dna microarray gene expression data distribution and associated moments. **Nature Biotechnology**, v. 11, n. 1, p. 1–14, Jan 2010.

TIMMINS, G. S.; DERETIC, V. Mechanisms of action of isoniazid. **Molecular Microbiology**, v. 62, n. 5, p. 1220–1227, Dez 2006.

TÖZEREN, A.; BYERS, S. W. **New Biology for Engineers and Computer Scientists**. Upper Saddle River, NJ: Pearson Prentice-Hall, 2004.

VARADAN, V. et al. The integration of biological pathway knowledge in cancer genomics. **IEEE Signal Processing Magazine**, v. 29, n. 1, p. 35–50, Jan 2012.

VARMA, A.; BOESCH, B. W.; PALSSON, B. O. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. **Appl Environ Microbiol**, v. 59, n. 8, p. 2465–73, Ago 1993.

VARMA, A.; PALSSON, B. O. Metabolic flux balancing: Basic concepts, scientific and practical use. **Nature Biotechnology**, Nature Publishing Group, v. 12, n. 10, p. 994–8, Jan 1994.

VILCHIÈZE, C.; JACOBS, W. R. The mechanism of isoniazid killing: clarity through the scope of genetics. **Annual Review of Microbiology**, v. 61, n. 1, p. 35–50, Jan 2007.

VOSKUIL, M. I. et al. Regulation of the *Mycobacterium tuberculosis* PE/PPE genes. **Tuberculosis**, v. 84, n. 3-4, p. 256–62, Jan 2004.

VOSS, J. J. D. et al. The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. **Proc Natl Acad Sci USA**, v. 97, n. 3, p. 1252–7, Fev 2000.

WANG, Z. et al. Stochastic dynamic modeling of short gene expression time-series data. **NanoBioscience, IEEE Transactions on**, v. 7, n. 1, p. 44–55, Mar 2008.

WEINKE, T. et al. Neuropsychiatric side effects after the use of mefloquine. **Am. J. Trop. Medic. and Hyg.**, v. 45, n. 1, p. 86–91, Jan 1991.

WIJAYA, E.; HARADA, H.; HORTON, P. Modeling the marginal distribution of gene expression with mixture models. In: **Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on**. [S.l.: s.n.], 2008. v. 3, p. 84–9.

WITTMAN, C. Fluxome analysis using GC-MS. **Microbial Cell Factories**, v. 6, n. 1, p. 6, Fev 2007.

WOOLEY, J.; LIN, H. **Catalyzing Inquiry at the Interface of Computing and Biology**. [S.l.], 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK25460/>.

YIZHAK, K. et al. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. **Bioinformatics**, v. 26, p. i255–i260, Jan 2010.

YOUNG, D. et al. Confronting the scientific obstacles to global control of tuberculosis. **The Journal of Clinical Investigation**, The American Society for Clinical Investigation, v. 118, n. 4, p. 1255–65, Abr 2008.

ZUPAN, B. et al. Genepath: a system for automated construction of genetic networks from mutant data. **Bioinformatics**, v. 19, n. 3, p. 383–9, Fev 2003.

ZUPKE, C.; SINSKEY, A. J.; STEPHANOPOULOS, G. Intracellular flux analysis applied to the effect of dissolved oxygen on hybridomas. **Appl Microbiol Biotechnol**, v. 44, n. 1, p. 27–36, Dez 1995.

APÊNDICE A – Mistura de Gamas Univariadas

Este apêndice mostra a obtenção das expressões recursivas para os parâmetros de uma mistura de PDFs gama, como utilizado no algoritmo do PathOlogist modificado.

As expressões apresentadas a seguir são as utilizadas na programação do método EM empregado neste trabalho.

Embora o método EM para estimação de mistura de distribuições gama esteja disponível em algumas linguagens (bem como a derivação das expressões (ALMHANA et al., 2006)), não havia uma versão disponível em R, fator importante para obter-se uma melhor integração do método com as outras simulações do trabalho.

Com essa maximização obtém-se expressões necessárias para o método EM para os dados de expressão genética modelados por uma mistura de duas distribuições gama.

Inicialmente, a função logaritmo da verossimilhança pode ser escrita como:

$$\ln[p(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\pi})] = \sum_{n=1}^{N} \ln\left[\sum_{k=1}^{K} \pi_k \mathcal{G}(x_n|\alpha_k,\beta_k)\right] \tag{A.1}$$

onde assume-se K um valor inteiro que identifica o número de componentes da mistura. Para o problema da expressão genética temos $K = 2$.

Como a expressão para a distribuição gama univariada é dada por:

$$\mathcal{G}(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}\mathrm{e}^{-\beta x} \tag{A.2}$$

Obtém-se para a expressão do logaritmo da verossimilhança:

$$\ln[p(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\pi})] = \sum_{n=1}^{N} \ln\left[\sum_{k=1}^{K} \pi_k \frac{1}{\Gamma(\alpha_k)}{\beta_k}^{\alpha_k}{x_n}^{\alpha_k-1}\mathrm{e}^{-\beta_k x_n}\right] \tag{A.3}$$

em que são assumidas as restrições $\alpha_k > 0$ e $\beta_k > 0$ para todo k, e que $\sum_{k=1}^{K} \pi_k = 1$.

O objetivo é encontrar o conjunto de parâmetros $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}$ que maximiza a expressão acima. Visto que não é possível encontrar expressões fechadas, eventualmente os valores deverão ser encontradas de forma recursiva com o uso do algoritmo EM (*expectation-maximization*).

Derivando em relação aos parâmetros $\boldsymbol{\beta_k}$:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta_k}} \ln[\boldsymbol{p}(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\pi})] = \sum_{n=1}^{N} \frac{\frac{\partial}{\partial\beta_k} \sum_{k=1}^{K} \boldsymbol{\pi_k} \frac{1}{\Gamma(\alpha_k)} \boldsymbol{\beta_k}^{\alpha_k} \boldsymbol{x_n}^{\alpha_k - 1} \mathrm{e}^{-\beta_k x_n}}{\sum_{j=1}^{K} \boldsymbol{\pi_j} \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_j},\boldsymbol{\beta_j})} \tag{A.4}$$

Obtém-se para cada $\boldsymbol{k}$ a expressão:

$$\sum_{n=1}^{N} \frac{\boldsymbol{\pi_k} \frac{1}{\Gamma(\alpha_k)} \boldsymbol{x_n^{\alpha_k - 1}} \left[\boldsymbol{\alpha_k} \boldsymbol{\beta_k}^{\alpha_k - 1} \mathrm{e}^{-\beta_k x_n} + \boldsymbol{\beta_k}^{\alpha_k} (-\boldsymbol{x_n}) \mathrm{e}^{-\beta_k x_n} \right]}{\sum_{j=1}^{K} \boldsymbol{\pi_j} \boldsymbol{\mathcal{G}} \dashv (\boldsymbol{x_n}|\boldsymbol{\alpha_j},\boldsymbol{\beta_j})} \tag{A.5}$$

Reorganizando os termos obtém-se duas PDFs gama no numerador:

$$\sum_{n=1}^{N} \frac{\boldsymbol{\pi_k} \left[\frac{\boldsymbol{\alpha_k}}{\boldsymbol{\beta_k}} \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_k},\boldsymbol{\beta_k}) + (-\boldsymbol{x_n}) \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_k},\boldsymbol{\beta_k}) \right]}{\sum_{j=1}^{K} \boldsymbol{\pi_j} \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_j},\boldsymbol{\beta_j})} \tag{A.6}$$

Utilizando-se a notação:

$$\boldsymbol{\gamma_{nk}} = \frac{\boldsymbol{\pi_k} \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_k},\boldsymbol{\beta_k})}{\sum_{j=1}^{K} \boldsymbol{\pi_j} \boldsymbol{\mathcal{G}}(\boldsymbol{x_n}|\boldsymbol{\alpha_j},\boldsymbol{\beta_j})} \tag{A.7}$$

Cada uma das derivadas parciais pode ser escrita como:

$$\sum_{n=1}^{N} \left(\frac{\boldsymbol{\alpha_k}}{\boldsymbol{\beta_k}} \boldsymbol{\gamma_{nk}} + (-\boldsymbol{x_n}) \boldsymbol{\gamma_{nk}} \right) \tag{A.8}$$

Igualando cada uma das derivadas a zero obtém-se a expressão desejada:

$$\frac{\boldsymbol{\alpha_k}}{\boldsymbol{\beta_k}} \sum_{n=1}^{N} \boldsymbol{\gamma_{nk}} = \sum_{n=1}^{N} \boldsymbol{x_n} \boldsymbol{\gamma_{nk}} \tag{A.9}$$

Usando a notação:

$$\boldsymbol{N_k} = \sum_{n=1} \boldsymbol{\gamma_{nk}} \tag{A.10}$$

Chega-se à expressão final para $\boldsymbol{\beta_k}$:

$$\boldsymbol{\beta_k} = \frac{\boldsymbol{\alpha_k} \boldsymbol{N_k}}{\sum_{n=1}^{N} \boldsymbol{x_n} \boldsymbol{\gamma_{nk}}} \tag{A.11}$$

Essa é a expressão para cada $\boldsymbol{\beta_k}$ que maximiza a verossimilhança. Porém, essa não é uma expressão fechada, visto que os valores $\boldsymbol{\gamma_{nk}}$ dependem de forma recursiva dos $\boldsymbol{\beta_k}$. A otimização deverá será feita de forma recursiva. Essa expressão é a mesma expressão apresentada em (EFRONI; SCHAEFER; BUETOW, 2007), apenas com a diferença que ali os autores definem $\boldsymbol{b = \frac{1}{\beta}}$ para a expressão da PDF gama.

A seguir apresentamos o cálculo das derivadas parciais em relação aos parâmetros $\boldsymbol{\alpha_k}$ e a obtenção das expressão ótimas que maximizam a verossimilhança. Derivando em relação aos $\boldsymbol{\alpha_k}$ tem-se:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial \alpha_k}} \ln[\boldsymbol{p}(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \alpha_k} \sum_{k=1}^{K} \pi_k \frac{1}{\Gamma(\alpha_k)} \beta_k{}^{\alpha_k} x_n{}^{\alpha_k - 1} \mathrm{e}^{-\beta_k x_n}}{\sum_{j=1}^{K} \pi_j \mathcal{G}(x_n|\alpha_j, \beta_j)} \tag{A.12}$$

E para cada $\boldsymbol{k}$ obtém-se:

$$\sum_{n=1}^{N} \frac{\pi_k \mathrm{e}^{-\beta_k x_n} \frac{\partial}{\partial \alpha_k} \left[\frac{1}{\Gamma(\alpha_k)} \beta_k{}^{\alpha_k} x_n{}^{\alpha_k} x_n^{-1} \right]}{\sum_{j=1}^{K} \pi_j \mathcal{G}(x_n|\alpha_j, \beta_j)} \tag{A.13}$$

Calculando-se separadamente a derivada parcial chega-se à expressão:

$$\begin{aligned} &\frac{\partial}{\partial \alpha_k} \frac{1}{\Gamma(\alpha_k)} (\beta_k x_n)^{\alpha_k} = \\ &- \frac{(\beta_k x_n)^{\alpha_k} \Psi(\alpha_k)}{\Gamma(\alpha_k)} + \frac{(\beta_k x_n)^{\alpha_k} \ln \alpha_k}{\Gamma(\alpha_k)} + \frac{\beta_k{}^{\alpha_k} \alpha_k x_n^{\alpha_k - 1}}{\Gamma(\alpha_k)} \end{aligned} \tag{A.14}$$

em que $\boldsymbol{\Psi(x)}$ é chamda *função digama*, i.e. a derivada do logaritmo da função gama. Para o resultado acima fez-se uso das propriedades:

(A).

$$\frac{d}{dx} a^x = a^x \ln a$$

para $\boldsymbol{a > 0}$, o que é possível, visto que $\boldsymbol{\beta_k > 0 \ \forall k}$;

(B).

$$\frac{d}{dx} \frac{1}{\Gamma(x)} = -\frac{\Psi(x)}{\Gamma(x)}$$

para $\boldsymbol{x > 0}$, o que é possível, visto que $\boldsymbol{\alpha_k > 0 \ \forall k}$.

em que a propriedade (B) é facilmente obtida observando-se que $\boldsymbol{\Gamma(x)^{-1} = \exp(-\ln(\Gamma(x)))}$.

Retornando o resultado da derivada na Eq. (A.13), coletando e reorganizando os termos chega-se, para cada $\boldsymbol{k}$, à expressão:

$$\sum_{n=1}^{N} \frac{\boldsymbol{\pi_k \mathcal{G}(x_n|\alpha_k, \beta_k)} \left[\boldsymbol{\alpha_k {x_n}^{-1} + \ln \alpha_k - \Psi(\alpha_k)}\right]}{\sum_{j=1}^{K} \boldsymbol{\pi_j \mathcal{G}(x_n|\alpha_j, \beta_j)}} \tag{A.15}$$

Utilizando-se novamente as notações compactas das Eqs. (A.7) e (A.10):

$$\sum_{n=1}^{N} \boldsymbol{\gamma_{nk}} \left(\frac{\boldsymbol{\alpha_k}}{\boldsymbol{x_n}} + \boldsymbol{\ln \alpha_k - \Psi(\alpha_k)}\right) \tag{A.16}$$

Igualando cada derivada a zero obtém-se a seguinte expressão para cada $\boldsymbol{\alpha_k}$:

$$\boldsymbol{\alpha_k} = \boldsymbol{N_k} \frac{\boldsymbol{\Psi(\alpha_k) - \ln(\alpha_k)}}{\sum_{n=1}^{N} \frac{\gamma_{nk}}{x_n}} \tag{A.17}$$

Essa expressão maximiza a verossimilhança, e mais uma vez a solução deve ser obtida de forma iterativa, visto que cada valor depende de $\boldsymbol{\alpha_k}$ através das funções $\boldsymbol{\Psi(\alpha_k)}$, $\boldsymbol{\Gamma(\alpha_k)}$ e $\boldsymbol{\gamma_{nk}}$.

Finalmente, para completar o processo de maximização ainda é necessário calcular as derivadas em relação aos $\boldsymbol{\pi_k}$ incluindo a restrição de que o seu somatório deve ser igual à unidade:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial \pi_k}} \boldsymbol{\ln[p(\mathrm{x}|\alpha, \beta, \pi)]} = \sum_{n=1}^{N} \frac{\boldsymbol{\mathcal{G}(x_n|\alpha_k, \beta_k)}}{\sum_{j=1}^{K} \boldsymbol{\pi_j \mathcal{G}(x_n|\alpha_j, \beta_j)}} + \frac{\boldsymbol{\partial}}{\boldsymbol{\partial \pi_k}} \boldsymbol{\lambda} \left(\sum_{k=1}^{K} \boldsymbol{\pi_k - 1}\right) \tag{A.18}$$

Igualando a derivada a zero para cada $\boldsymbol{k}$:

$$-\boldsymbol{\lambda} = \sum_{n=1}^{N} \frac{\boldsymbol{\mathcal{G}(x_n|\alpha_k, \beta_k)}}{\sum_{j=1}^{K} \boldsymbol{\pi_j \mathcal{G}(x_n|\alpha_j, \beta_j)}} \tag{A.19}$$

Multiplicando ambos os lados por $\boldsymbol{\pi_k}$:

$$-\boldsymbol{\pi_k \lambda} = \sum_{n=1}^{N} \frac{\boldsymbol{\pi_k \mathcal{G}(x_n|\alpha_k, \beta_k)}}{\sum_{j=1}^{K} \boldsymbol{\pi_j \mathcal{G}(x_n|\alpha_k, \beta_k)}} = \sum_{n=1}^{N} \boldsymbol{\gamma_{nk}} \tag{A.20}$$

Somando para todo $\boldsymbol{k = 1 \ldots K}$:

$$-\boldsymbol{\lambda} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \mathbf{1} \quad \Rightarrow \quad \boldsymbol{\lambda} = -\boldsymbol{N} \tag{A.21}$$

Usando novamente as formas compactas (A.7) e (A.10):

$$\boldsymbol{\pi_k} = \frac{\boldsymbol{N_k}}{\boldsymbol{N}} \tag{A.22}$$

Essas expressões também são recursivas, visto que $\boldsymbol{N_k}$ é função de $\boldsymbol{\gamma_{nk}}$, a qual por sua vez é função de $\boldsymbol{\pi_k}$.

APÊNDICE B – Apresentação do modelo

Desenvolvimento analítico detalhado do modelo de regressão RVM apresentado em (CHAKRABORTY; GHOSH; MALLICK, 2012).

B.1 MODELO DE DADOS

Propõe-se estimar fluxos metabólicos a partir de dados de expressão genética com um modelo de regressão multivariável com variáveis latentes. O modelo de dados é:

$$\boldsymbol{v_i} = \boldsymbol{z_i} + \boldsymbol{\eta_i} \tag{B.1}$$

$$\boldsymbol{z_i} = \boldsymbol{K_i^0}\boldsymbol{\beta^0} + \boldsymbol{\delta_i} \tag{B.2}$$

em que:

- $\boldsymbol{i = 1 \ldots n}$ indica o número de observações no conjunto de dados $\boldsymbol{\mathcal{D}\{e_i, v_i\}}$

- assume-se que o vetor $\boldsymbol{\eta_i} = (\boldsymbol{\eta_{i1}, \ldots, \eta_{iq}})^{\boldsymbol{T}}$ é uma variável aleatória multivariada de dimensão $\boldsymbol{q}$, com elementos independentes e normalmente distribuídos de média zero e matriz de covariância $\boldsymbol{\Sigma_\eta} = \text{diag}(\boldsymbol{\sigma_1^2, \ldots, \sigma_q^2})$

$$\boldsymbol{p(\eta|\Sigma_\eta)} = \frac{\boldsymbol{1}}{\boldsymbol{(2\pi)^{q/2}}} \frac{\boldsymbol{1}}{\boldsymbol{|\Sigma_\eta|^{1/2}}} \boldsymbol{\exp}\left(-\frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\eta^T \Sigma_\eta^{-1} \eta}\right) \tag{B.3}$$

- o vetor de erros residuais $\boldsymbol{\delta_i} = (\boldsymbol{\delta_{i1}, \ldots, \delta_{iq}})^{\boldsymbol{T}}$ é uma variável aleatória multivariada de dimensão $\boldsymbol{q}$ com elementos independentes e normalmente distribuídos de média zero e matriz de covariância $\boldsymbol{\Sigma_\delta}$. Essa variável, junto com a variável $\boldsymbol{\eta}$, modela toda a fonte de aleatoriedade que não é explicada pelo modelo de regressão.

$$\boldsymbol{p(\delta|\Sigma_\delta)} = \frac{\boldsymbol{1}}{\boldsymbol{(2\pi)^{q/2}}} \frac{\boldsymbol{1}}{\boldsymbol{|\Sigma_\delta|^{1/2}}} \boldsymbol{\exp}\left(-\frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\delta^T \Sigma_\delta^{-1} \delta}\right) \tag{B.4}$$

- a matriz $\boldsymbol{K_i^0}$ é uma matriz bloco-diagonal com cada bloco igual ao vetor $\boldsymbol{K_i} = (\boldsymbol{1, k(e_i, e_1|\theta), \ldots, k(e_i, e_n|\theta)})$ e definida como $\boldsymbol{K_i^0} = \boldsymbol{I_q} \otimes \boldsymbol{K_i}$, em que $\boldsymbol{I_q}$ é a matriz identidade de dimensão $\boldsymbol{q}$, o símbolo $\otimes$ indica a operação produto de Kronecker, e $\boldsymbol{\theta}$ é parâmetro do kernel $\boldsymbol{k(\cdot, \cdot)}$. Essa matriz define as funções-base

não-lineares do modelo de regressão.

$$\boldsymbol{K_i^0} = \begin{bmatrix} \boldsymbol{K_i} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K_i} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{K_i} \end{bmatrix}_{\boldsymbol{q \times (n+1)q}} \tag{B.5}$$

- o vetor de coeficientes de regressão $\boldsymbol{\beta^0}$ é um vetor coluna de dimensão $\boldsymbol{q(n+1)}$ composto pelos $\boldsymbol{q}$ vetores de regressão para cada elemento $\boldsymbol{j}$ do vetor $\boldsymbol{v}$, definido como $\boldsymbol{\beta^0 = (\beta_1^T, \beta_2^T, \dots, \beta_q^T)^T}$:

$$\boldsymbol{\beta^0 = (\beta_{01}, \dots, \beta_{n1}, \beta_{02}, \beta_{12}, \dots, \beta_{n2}, \dots, \beta_{0q}, \beta_{1q}, \dots, \beta_{nq})^T} \tag{B.6}$$

- $\boldsymbol{z_i}$ são variáveis latentes introduzidas no modelo:

$$\boldsymbol{p(z_i|K_i^0\beta^0, \Sigma_\delta) = \frac{1}{(2\pi)^{q/2}} \frac{1}{|\Sigma_\delta|^{1/2}} \exp\left(-\frac{1}{2}(z_i - K_i^0\beta^0)^T \Sigma_\delta^{-1}(z_i - K_i^0\beta^0)\right)} \tag{B.7}$$

Cada elemento $\boldsymbol{z_{ij}}$ do vetor $\boldsymbol{z_i}$ é dado por:

$$\boldsymbol{z_{ij} = \beta_{0j} + \beta_{1j}k(e_i, e_1) + \beta_{2j}k(e_i, e_2) + \dots + \beta_{nj}k(e_i, e_n) + \delta_{ij}}$$

que pode ser escrita como $\boldsymbol{z_{ij} = K_i\beta_j + \delta_{ij}}$. É fácil verificar que $\boldsymbol{z_i = K_i^0\beta + \delta_i}$, conforme indicado no modelo de dados (B.2).

Esse modelo de regressão define uma distribuição coondicional para as observações de saída $\boldsymbol{v_i}$ na forma:

$$\boldsymbol{p(v|z, \Sigma_\eta) = \mathcal{N}(z, \Sigma_\eta)} \tag{B.8}$$

em que a média é dada pelo vetor de variáveis latentes $\boldsymbol{z}$ com distribuição(CHAKRABORTY; GHOSH; MALLICK, 2012):

$$\boldsymbol{p(z|e, \Theta_z) = \mathcal{N}(f(e), \Sigma_\delta)} \tag{B.9}$$

em que $\boldsymbol{\Theta_z}$ indica o conjunto de parâmetros usados na Eq. (B.2) do modelo de dados.

B.2 FUNÇÃO VEROSSIMILHANÇA

A função verossimilhança para o conjunto dos n dados observados e das variáveis latentes é dada por(BISHOP, 2007):

$$p(V, Z|\Theta) = p(V|Z, \Theta)p(Z|\Theta) \tag{B.10}$$

em que $V = (v_1, v_2, \ldots, v_n)$ e $Z = (z_1, z_2, \ldots, z_n)$ representam respectivamente o conjunto dos n vetores v_i observados e respectivos vetores latentes z_i. Assumindo os vetores v e z i.i.d., segue que as distribuições do lado direito da equação (B.10) serão dadas pelo produtório das n distribuições gaussianas multivariadas:

$$p(v_i|z_i, \Theta) \sim \mathcal{N}(z_i, \Sigma_\eta) \tag{B.11}$$

$$p(z_i|\Theta) \sim \mathcal{N}(K_i^0\beta, \Sigma_\delta) \tag{B.12}$$

Resultando nas expressões:

$$p(V|Z, \Theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{q/2}} \frac{1}{|\Sigma_\eta|^{1/2}} \exp\left(-\frac{1}{2}(v_i - z_i)^T \Sigma_\eta^{-1}(v_i - z_i)\right) \tag{B.13}$$

e:

$$p(Z|\Theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{q/2}} \frac{1}{|\Sigma_\delta|^{1/2}} \exp\left(-\frac{1}{2}(z_i - K_i^0\beta^0)^T \Sigma_\delta^{-1}(z_i - K_i^0\beta^0)\right) \tag{B.14}$$

B.3 DISTRIBUIÇÕES *A PRIORI*

Para os parâmetros do modelo em Θ definem-se as distribuições *a priori* a seguir baseadas na propriedade da conjugação, exceto para o parâmetro do kernel, onde optou-se pelo prior uniforme.

Para o kernel polinomial define-se uma distribuição *a priori* uniforme discreta:

$$p(\theta) = \mathcal{U}\{1, 2, \ldots, C\} \tag{B.15}$$

em que a probabilidade de cada valor possível para a variável é igual a

$\mathbf{1/C}$. Para o kernel gaussiano utiliza-se um prior uniforme contínuo:

$$p(\theta) = \mathcal{U}(\theta_{min}, \theta_{max}) \tag{B.16}$$

No caso das máquinas de vetor de relevância, assume-se que os coeficientes de regressão em $\boldsymbol{\beta^0}$ são independentes e normalmente distribuídos(SCHÖLKOPF; SMOLA, 2002). Cada um dos $\boldsymbol{q}$ vetores de regressão $\boldsymbol{\beta_j}$ (i.e. coeficientes da regressão para cada elemento $\boldsymbol{v_j}$) recebe como prior uma distribuição normal com média zero e matriz de covariância $\boldsymbol{\Lambda_j^{-1}}$:

$$p(\beta_j|\Lambda_j) = \mathcal{N}_{n+1}(0, \Lambda_j^{-1}) \tag{B.17}$$

em que $\boldsymbol{j = 1, \ldots, q}$ e $\boldsymbol{\Lambda_j} = \text{diag}(\boldsymbol{\lambda_{0j}, \lambda_{1j}, \ldots, \lambda_{nj}})$.

Para o vetor completo de todos os coeficientes de regressão $\boldsymbol{\beta^0}$ definido na Eq. (B.6), o prior é dado por:

$$p(\beta^0|\Lambda) = \frac{1}{(2\pi)^{q(n+1)/2}} \frac{1}{|\Lambda|^{-1/2}} \exp\left(-\frac{1}{2}\beta^{0T}\Lambda\beta^0\right) \tag{B.18}$$

em que a matriz $\boldsymbol{\Lambda}$ é bloco-diagonal:

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \Lambda_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \Lambda_q \end{bmatrix}_{[(n+1)q]\times[(n+1)q]} \tag{B.19}$$

Cada um dos hiperparâmetros $\boldsymbol{\lambda_{ij}}$ representa um valor de precisão (inverso da variância) e recebe um prior gama:

$$p(\lambda_{ij}) = \frac{1}{\Gamma(c)} d^c \lambda_{ij}^{c-1} \exp(-d\lambda_{ij}) \tag{B.20}$$

sendo que os parâmetros $\boldsymbol{c}$ e $\boldsymbol{d}$ devem ser definidos de acordo com a forma desejada para o prior. Esses valores, para um prior não-informativo em geral são tomados como iguais a $\mathbf{10^{-4}}$, como indicado em (SCHÖLKOPF; SMOLA, 2002). É o uso de hiperparâmetros $\boldsymbol{\lambda_{ij}}$ individuais para cada coeficiente de regressão o que permite esparsidade (e consequentemente baixa complexidade do modelo e suavidade da função aprendida) no modelo RVM. Durante o treinamento, os hiperparâmetros $\boldsymbol{\lambda_{ij}}$ correspondentes aos coeficientes de regressão $\boldsymbol{\beta_{ij}}$ que não contribuem para maximização da função de verossimilhança (i.e.

não contribuem para a função mais provável de ter gerado os dados observados), acabam sendo maximizados para valores muito grandes, efetivamente eliminando os coeficientes correspondentes do modelo de regressão ((BISHOP, 2007),cap. 7).

No caso do vetor de efeitos aleatórios residuais $\boldsymbol{\delta}$ assumiu-se a distribuição normal da Eq. (B.4), e considera-se um prior wishart inverso para a matriz de covariância $\boldsymbol{\Sigma_\delta}$:

$$\boldsymbol{p(\Sigma_\delta)} = \frac{\boldsymbol{|Q|^{\frac{\nu}{2}}}}{\boldsymbol{2^{\frac{\nu q}{2}}\Gamma_q(\frac{\nu}{2})}} \boldsymbol{|\Sigma_\delta|^{-0.5(\nu+q+1)}} \boldsymbol{\exp}\left(-\frac{\boldsymbol{1}}{\boldsymbol{2}}\mathrm{tr}(\boldsymbol{Q\Sigma_\delta^{-1}})\right) \quad \text{(B.21)}$$

em que $\boldsymbol{\nu}$ são os graus de liberdade da distribuição e $\boldsymbol{Q}$ é a matriz de escala. Para que a distribuição seja própria, é necessário que o número de graus de liberdade seja no mínimo igual à dimensão do vetor $\boldsymbol{\delta_i}$. A matriz de escala deve ser simétrica e positiva definida, como por exemplo a matriz identidade.

Visto que assume-se uma distribuição normal multivariável com matriz de covariância diagonal para o vetor $\boldsymbol{\eta}$, define-se separadamente a distribuição prior para cada uma das variâncias $\boldsymbol{\sigma_{\eta_1}, \ldots, \sigma_{\eta_q}}$ com uma função gama inversa:

$$\boldsymbol{p(\sigma^2_{\eta_j})} = \frac{\boldsymbol{\gamma_2^{\gamma_1}}}{\boldsymbol{\Gamma(\gamma_1)}} \boldsymbol{(\sigma^2_{\eta_j})^{-\gamma_1-1}} \boldsymbol{\exp}\left(-\frac{\boldsymbol{\gamma_2}}{\boldsymbol{\sigma^2_{\eta_j}}}\right) \quad \text{(B.22)}$$

B.4 DISTRIBUIÇÃO *A POSTERIORI*

A distribuição posterior conjunta para os parâmetros do modelo é dada pela multiplicação dos priors e da função verossimilhança.

Assumindo que os vetores $\boldsymbol{v_i}$ são independentes condicionados em $\boldsymbol{z_i}$, e descartando os termos de normalização de cada distribuição, é possível escrever a expressão para a distribuição conjunta *a posteriori* dos parâmetros do modelo $\boldsymbol{p(Z, \beta^0, \Sigma_\eta, \Sigma_\delta, \Lambda, \theta|V)}$ como sendo proporcional à multiplicação dos seguintes termos[1]:

•verossimilhança para os dados $\boldsymbol{v}$:

$$\boldsymbol{\exp}\left(-\frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{(v_i - z_i)^T\Sigma_\eta^{-1}(v_i - z_i)}\right) \quad \text{(B.23)}$$

[1] cf. (CHAKRABORTY; GHOSH; MALLICK, 2012)

•verossimilhança da variável latente $\boldsymbol{z}$:

$$\frac{\mathbf{1}}{|\boldsymbol{\Sigma_\delta}|^{n/2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\sum_{i=1}^{n}(z_i - \boldsymbol{K}_i^0\boldsymbol{\beta}^0)^T\boldsymbol{\Sigma}_\delta^{-1}(z_i - \boldsymbol{K}_i^0\boldsymbol{\beta}^0)\right) \tag{B.24}$$

•prior para a matriz de covariância $\boldsymbol{\Sigma_\delta}$:

$$|\boldsymbol{\Sigma_\delta}|^{-0.5(\nu+q+1)} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}_\delta^{-1})\right) \tag{B.25}$$

•prior para a matriz de covariância $\boldsymbol{\Sigma_\eta}$:

$$\prod_{j=1}^{q} \exp\left(-\frac{\gamma_2}{\sigma_j^2}\right)(\sigma_j^2)^{-\gamma_1-1} \tag{B.26}$$

•prior para a matriz de coeficientes de regressão $\boldsymbol{\beta}$:

$$\frac{\mathbf{1}}{|\boldsymbol{\Lambda}|^{-1/2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\beta}^{0T}\boldsymbol{\Lambda}\boldsymbol{\beta}^0\right) \tag{B.27}$$

•prior para a matriz de precisão $\boldsymbol{\Lambda}$:

$$\prod_{i=1}^{n}\prod_{j=1}^{q} \lambda_{ij}^{c-1} \exp(-d\lambda_{ij}) \tag{B.28}$$

As distribuições *a priori* para o parâmetro $\boldsymbol{\theta}$ podem ser descartadas, visto que são apenas valores constantes que independem do valor da variável aleatória.

As distribuições de verossimilhança na forma de produtórios foram reescritas na forma de somatórios no argumento da função exponencial.

B.5 AMOSTRAGEM MCMC

Como a distribuição *a posteriori* conjunta definida pelas Eqs. (B.23) a (B.28) é demasiada complexa para permitir uma solução analítica, serão aplicados os métodos de amostragem Markov Chain Monte Carlo para obter uma estimativa da distribuição. O amostrador de Gibbs (ROBERT; CASELLA, 2004) amostra de PDFs condicionais de cada parâmetro condicionado em todas outras variáveis. A seguir são apresentadas as

expressões para cada uma das PDFs condicionais.

B.5.1 Expressão para $p(\beta^0|\Lambda, Z, V, \Sigma_\eta, \Sigma_\delta, \theta)$

A partir das Eqs. (B.24) e (B.27), descartando todos os termos que não dependem de $\boldsymbol{\beta^0}$ e portanto são constantes, podemos escrever:

$$\boldsymbol{p_{\beta^0}} \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta^{0T}\Lambda\beta^0}\right)\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{z_i - K_i^0\beta^0})^T\boldsymbol{\Sigma_\delta^{-1}}(\boldsymbol{z_i - K_i^0\beta^0})\right) \tag{B.29}$$

em que $\boldsymbol{p_{\beta^0}}$ indica a PDF do vetor de coeficientes de regressão $\boldsymbol{\beta^0}$ condicionada em todos os outros parâmetros e na variável latente e nos dados de saída.

Combinando as duas funções exponenciais temos:

$$\boldsymbol{p_{\beta^0}} \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta^{0T}\Lambda\beta^0} + \sum_{i=1}^{n}(\boldsymbol{z_i - K_i^0\beta^0})^T\boldsymbol{\Sigma_\delta^{-1}}(\boldsymbol{z_i - K_i^0\beta^0})\right)\right) \tag{B.30}$$

Desenvolvendo o argumento da função exponencial temos:

$$\begin{aligned}\ln \boldsymbol{p_{\beta^0}} \propto -\frac{1}{2}\Big(&\boldsymbol{\beta^{0T}\Lambda\beta^0} + \\ &\sum_{i=1}^{n}\left[\boldsymbol{z_i^T\Sigma_\delta^{-1}z_i} - (\boldsymbol{K_i^0\beta^0})^T\boldsymbol{\Sigma_\delta^{-1}z_i} - \boldsymbol{z_i^T\Sigma_\delta^{-1}K_i^0\beta^0}\right] + \\ &\boldsymbol{\beta^{0T}}(\sum_{i=1}^{n}\boldsymbol{K_i^{0T}\Sigma_\delta^{-1}K_i^0})\boldsymbol{\beta^0}\Big)\end{aligned} \tag{B.31}$$

Como:

$$\begin{aligned}\boldsymbol{z_i^T\Sigma_\delta^{-1}K_i^0\beta^0} = (\boldsymbol{\Sigma_\delta^{-1}z_i})^T\boldsymbol{K_i^0\beta^0} = \\ (\boldsymbol{K_i^0\beta^0})^T\boldsymbol{\Sigma_\delta^{-1}z_i} = \boldsymbol{\beta^{0T}K_i^{0T}\Sigma_\delta^{-1}z_i}\end{aligned} \tag{B.32}$$

Segue que:

$$\ln p_{\beta^0} \propto -\frac{1}{2}\beta^{0T}\Big(\sum_{i=1}^{n} K_i^{0T}\Sigma_\delta^{-1}K_i^0 + \Lambda\Big)\beta^0 + \beta^{0T}\sum_{i=1}^{n} K_i^{0T}\Sigma_\delta^{-1}z_i + \sum_{i=1}^{n} z_i^T\Sigma_\delta^{-1}z_i \quad \text{(B.33)}$$

Comparando esta últma expressão com a expressão correspondente para a distribuição gaussiana genérica:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^t\Sigma^{-1}\mu \quad \text{(B.34)}$$

Temos que a PDF p_{β^0} é normal multivariada com vetor de média μ_{β^0} e matriz de covariância Σ_{β^0} iguais a:

$$\mu_{\beta^0} = \Sigma_{\beta^0}\left(\sum_{i=1}^{n} K_i^{0T}\Sigma_\delta^{-1}z_i\right) \quad \text{(B.35)}$$

$$\Sigma_{\beta^0} = \left(\sum_{i=1}^{n} K_i^{0T}\Sigma_\delta^{-1}K_i^0 + \Lambda\right)^{-1} \quad \text{(B.36)}$$

B.5.2 Expressão para $p(\Sigma_\delta|\beta^0, \Lambda, Z, V, \Sigma_\eta, \theta)$

A partir das Eqs. (B.24) e (B.25), descartando todos os termos que não dependem de Σ_δ e portanto são constantes, podemos escrever:

$$p_{\Sigma_\delta} \propto \frac{1}{|\Sigma_\delta|^{n/2}}|\Sigma_\delta|^{-0.5(\nu+q+1)} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}(Q\Sigma_\delta^{-1}) - \frac{1}{2}\sum_{i=1}^{n}(z_i - K_i^0\beta^0)^T\Sigma_\delta^{-1}(z_i - K_i^0\beta^0)\right) \quad \text{(B.37)}$$

em que p_{Σ_δ} indica a PDF da matriz de covariância Σ_δ condicionada em todos os outros parâmetros e na variável latente e nos dados de saída.

A partir da propriedade cíclica do traço e observando que $\det(A\cdot$

$\boldsymbol{B)} = \mathbf{det}(\boldsymbol{A})\,\mathbf{det}(\boldsymbol{B})$ é possível escrever:

$$p_{\boldsymbol{\Sigma_\delta}} \propto |\boldsymbol{\Sigma_\delta}|^{-0.5(n+\nu+q+1)} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma_\delta^{-1}}) - \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}(\boldsymbol{z_i} - \boldsymbol{K_i^0}\boldsymbol{\beta^0})(\boldsymbol{z_i} - \boldsymbol{K_i^0}\boldsymbol{\beta^0})^T\boldsymbol{\Sigma_\delta^{-1}}\right)\right) \tag{B.38}$$

Usando a notação $\boldsymbol{S_{zn}}$ para a matriz de covariância amostral de $\boldsymbol{z}$ multiplicada por $\boldsymbol{n-1}$, escrevemos:

$$p_{\boldsymbol{\Sigma_\delta}} \propto |\boldsymbol{\Sigma_\delta}|^{-0.5(n+\nu+q+1)} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma_\delta^{-1}}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{S_{zn}}\boldsymbol{\Sigma_\delta^{-1}})\right) \tag{B.39}$$

Finalmente, devido à propriedade do traço $\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$ podemos escrever:

$$p_{\boldsymbol{\Sigma_\delta}} \propto |\boldsymbol{\Sigma_\delta}|^{-0.5(n+\nu+q+1)} \exp\left(-\frac{1}{2}\mathrm{tr}\left[(\boldsymbol{Q} + \boldsymbol{S_{zn}})\boldsymbol{\Sigma_\delta^{-1}}\right]\right) \tag{B.40}$$

Comparando esta última expressão com a forma genérica para a distribuição Wishart inversa, identificam-se os parâmetros matriz de escala $\boldsymbol{Q_{\Sigma_\delta}}$ e graus de liberdade $\boldsymbol{\nu_{\Sigma_\delta}}$ abaixo:

$$\boldsymbol{Q_{\Sigma_\delta}} = \boldsymbol{Q} + \boldsymbol{S_{zn}} = \boldsymbol{Q} + \sum_{i=1}^{n}(\boldsymbol{z_i} - \boldsymbol{K_i^0}\boldsymbol{\beta^0})(\boldsymbol{z_i} - \boldsymbol{K_i^0}\boldsymbol{\beta^0})^T \tag{B.41}$$

$$\boldsymbol{\nu_{\Sigma_\delta}} = \boldsymbol{n} + \boldsymbol{\nu} \tag{B.42}$$

B.5.3 Expressão para $p(\sigma^2_{\eta_j}|\beta^0, \Lambda, Z, V, \Sigma_\delta, \theta)$

A partir das Eqs. (B.23) e (B.26), descartando todos os termos que não dependem de Σ_η e portanto são constantes, podemos escrever:

$$p_{\sigma^2_{\eta_j}} \propto (\sigma^2_{\eta_j})^{-\gamma_1-1} \exp\left(-\frac{\gamma_2}{\sigma^2_{\eta_j}}\right) \frac{1}{(\sigma^2_{\eta_j})^{n/2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(v_i - z_i)^T \Sigma_\eta^{-1}(v_i - z_i)\right) \quad \text{(B.43)}$$

em que $p_{\sigma^2_{\eta_j}}$ indica a PDF de cada elemento da diagonal da matriz de covariância Σ_δ condicionada em todos os outros parâmetros e na variável latente e nos dados de saída.

Como no somatório apenas a variável com índice j é variável, é possível escrever:

$$p_{\sigma^2_{\eta_j}} \propto (\sigma^2_{\eta_j})^{-\gamma_1-1-n/2} \exp\left(-\frac{1}{\sigma^2_{\eta_j}}\left(\gamma_2 + \frac{1}{2}\sum_{i=1}^{n}(v_{ij} - z_{ij})^2\right)\right) \quad \text{(B.44)}$$

Comparando esta última expressão com a expressão para a distribuição gama inversa, podemos identificar os parâmetros:

$$\gamma_{1_{\sigma^2_j}} = \gamma_1 + \frac{n}{2} \quad \text{(B.45)}$$

$$\gamma_{2_{\sigma^2_j}} = \gamma_2 + \frac{\sum_{i=1}^{n}(v_{ij} - z_{ij})^2}{2} \quad \text{(B.46)}$$

B.5.4 Expressão para $p(\lambda_{ij}|\beta^0, Z, V, \Sigma_\delta, \Sigma_\eta, \theta)$

A partir das Eqs. (B.27) e (B.28), descartando todos os termos que não dependem de λ_{ij} e portanto são constantes, podemos escrever:

$$p_{\lambda_{ij}} \propto \frac{1}{\lambda_{ij}^{-1/2}} \exp\left(-\frac{1}{2}\beta_{ij}\lambda_{ij}\right) \frac{1}{\Gamma(c)} d^c \lambda_{ij}^{c-1} \exp(-d\lambda_{ij}) \quad \text{(B.47)}$$

em que $p_{\lambda_{ij}}$ indica a PDF de cada elemento da diagonal da matriz Λ condicionada em todos os outros parâmetros, na variável latente e nos

dados de saída.

Combinando termos similares e simplificando:

$$p_{\lambda_{ij}} \propto \lambda_{ij}^{(c+1/2)-1} \exp\left(-\left(\frac{\beta_{ij}}{2}+d\right)\lambda_{ij}\right) \tag{B.48}$$

Finalmente, comparando esta última expressão com a expressão para a distribuição gama, podemos identificar os parâmetros:

$$c_\lambda = c + \frac{1}{2} \tag{B.49}$$

$$d_\lambda = \frac{\beta_{ij}}{2} + d \tag{B.50}$$

B.5.5 Expressão para $p(z_i|\beta^0, \Lambda, V, \Sigma_\delta, \Sigma_\eta, \theta)$

A partir das Eqs. (B.23) e (B.27), descartando todos os termos que não dependem de z_i e portanto são constantes, podemos escrever:

$$p_{z_i} \propto \exp\Big(-\frac{1}{2}\big((v_i - z_i)^T\Sigma_\eta^{-1}(v_i - z_i) + (z_i - K_i^0\beta^0)^T\Sigma_\delta^{-1}(z_i - K_i^0\beta^0)\big)\Big) \tag{B.51}$$

em que p_{z_i} indica a PDF de cada vetor z_i condicionada em todos os outros parâmetros e nos dados de saída.

Seguindo um procedimento similar ao usado na obtenção de $p(\beta^0|\Lambda, Z, V, \Sigma_\eta, \Sigma_\delta, \theta)$, identificamos p_{z_i} como uma distribuição normal multivariável com parâmetros média e matriz de covariância dados por:

$$\mu_{z_i} = \Sigma_{z_i}(\Sigma_\eta^{-1}v_i + \Sigma_\delta^{-1}K_i^0\beta^0) \tag{B.52}$$

$$\Sigma_{z_i} = (\Sigma_\eta^{-1} + \Sigma_\delta^{-1})^{-1} \tag{B.53}$$

B.5.6 Expressão para $p(\theta|\beta^0, \Lambda, V, Z, \Sigma_\delta, \Sigma_\eta)$

Observando a Eq. (B.27) vemos que o parâmetro θ entra na expressão como parâmetro do kernel $k(\cdot,\cdot)$ que perfaz cada elemento da matriz K_i^0 e portanto apresenta uma distribuição não trivial. A amos-

tragem dessa PDF deverá ser feita com um método do tipo aceita-rejeita, a partir de uma distribuição proposta da qual é possível amostrar (e.g. gaussiana).

B.6 FUNÇÕES KERNEL

Diferentes funções kernel podem ser utilizadas para construir a matriz $\boldsymbol{K_i^0}$ a partir dos vetores de expressão genética. Inicialmente propõe-se analisar este modelo para o kernel gaussiano:

$$\boldsymbol{K(e_i, e_j|\theta) = \exp\left(-\frac{\|e_i - e_j\|^2}{2\theta}\right)} \tag{B.54}$$

e o kernel polinomial:

$$\boldsymbol{K(e_i, e_j|\theta) = (e_i^T e_j + 1.0)^\theta} \tag{B.55}$$

No caso do kernel (B.54) o parâmetro $\boldsymbol{\theta}$ é a largura de banda do kernel, um indicativo da capacidade do kernel de discernir a similaridade entre os dois vetores do argumento. Para o kernel polinomial o parâmetro $\boldsymbol{\theta}$ é o grau do polinômio, e indica quais monômios são usados como descritores no espaço das características.

APÊNDICE C – Glossário

As definições a seguir fornecem explicações para fundamentos e conceitos importantes da biologia necessários para uma primeira leitura do trabalho. Os conceitos são explicados de forma simplificada e informações mais detalhadas podem ser obtidas, por exemplo, em (TÖZEREN; BYERS, 2004) ou (CAMPBELL; REECE, 2009).

ácido graxo: molécula orgânica produzida quando gorduras são quebradas. Sob certas condições, os ácidos graxos podem ser utilizados como fonte de energia pela célula.

alinhamento: processo de comparação de duas sequências de proteínas, DNA ou RNA com o objetivo de encontrar regiões de similaridade dentro das duas sequências.

alvo: na pesquisa e desenvolvimento de antibióticos, um alvo pode ser um gene, uma proteína ou uma reação química do organismo patogênico que tem a sua função alterada pela ação de um composto antibiótico, causando a morte do organismo.

amino-ácido: compostos orgânicos importantes nos organismos vivos. Um subconjunto de 20 amino-ácidos é utilizado por todos organismos para a síntese de proteínas.

amplo espectro, composto de: composto de ação antibiótica não específica, atuando sobre múltiplos alvos.

anabolismo: metabolismo construtivo. Em organismos vivos, é a síntese de moléculas mais complexas a partir de moléculas mais simples (oposto a catabolismo).

anoxia: ausência de oxigênio.

apoptose: morte programada da célula.

atividade bactericida: atividade que mata ou produz morte de bactérias.

biomassa: massa total de um organismo (e.g. célula bacterial).

biossíntese: produção de um molécula complexa em organismos vivos.

bomba de efluxo: proteína de transporte situadas na membrana celular, que pode apresentar ação anti-bactericida, retirando compostos tóxicos de dentro da célula e movendo-os para o ambiente extracelular.

caminho metabólico: subconjunto das reações químicas que compõem o metabolismo de um organismo e que atuam em conjunto para realização de uma atividade metabólica específica.

catabolismo: metabolismo destrutivo. Em organismos vivos, é a quebra de substâncias mais complexas em outras mais simples.

catálise: aumento da velocidade de uma reação química, por exemplo, quando em presença de uma enzima catalisadora.

cepa resistente: linhagem de células bacteriais com mutações que as tornam resistentes a compostos antibióticos.

cinética enzimática: estudo da velocidade das reações químicas catalisadas por enzimas.

endergônica: reação química que absorve energia.

enzima: proteínas complexas que catalizam atividades celulares específicas.

estequiometria: relaciona quantitativamente as quantidades de reagentes e produtos em uma reação química.

eucariótica: célula com um núcleo definido.

exergônica: reação química que produz energia.

fator de transcrição: proteínas regulatórias que regulam a expressão de outros genes e elas mesmas através da ativação ou inibição do processo de transcrição. Esses fatores de transcrição são os que atuam na ativação de um novo programa de expressão quando a célula enfrenta uma nova condição ambiental (poucos nutrientes, doença, drogasi antibióticas, etc.).

fenótipo: característica bioquímica aparente ou não, resultante da interação de um genótipo com o ambiente, adicionado à uma variação aleatória.

fenótipo KO: fenótipos de nocaute (um fenótipo que ocorre quando um gene é nocauteado).

fisiologia: o estudo do funcionaimento de um organismo vivo e suas partes.

fluxo metabólico: taxa de produção ou consumo de um composto químico em um caminho metabólico.

fosfolipídio: lípidio que contém grupos fosfato na sua composição. São importantes componentes na constituição da parede celular.

fosforilação: adição de um grupo fosfato (PO_4^{3-}) a uma molécula.

gene: uma curta sequência de DNA responsável por aspectos funcionais e estruturais de uma célula. Todo organismo contém um grande número de genes que são responsáveis por características (aparentes ou não) desse organismo, tais como cor do olho e tipo sanguíneo (fenótipos).

genoma: o DNA completo de um organismo.

genômica: o estudo dos genomas para identificação e sequenciamento de genes, e a sua aplicação.

hidrólise: quebra da ligação de uma molécula ao reagir com uma molécula de água.

hipoxia: baixa concentração de oxigênio.

homeostase: equilíbrio de um processo ou de um conjunto de processos fisiológicos interconectados.

homologia: similaridades de função entre estruturas de organismos de diferentes espécies (não apenas espécies descendentes de um mesmo ancestral).

***in silico*:** técnica ou procedimento realizada em um computador ou por uma simulação computacional.

***in vitro*:** técnica experimental realizada em um ambiente controlado fora de um organismo.

***in vivo*:** técnica experimental realizada com um organismo vivo completo.

inibidor da respiração: composto antibiótico que interfere com o processo de respiração celular.

inibidor da síntese da parede celular: composto antibiótico que impede o organismo de sintetizar as moléculas constituintes da parede celular.

inibidor de síntese protéica: composto antibiótico que interfere com o processo de tradução.

inibidor transcripcional: composto antibiótico que interfere com o processo de transcrição.

interactoma: o conjunto de todas as interações existentes entre proteínas dentro de uma célula.

lipoproteína: complexo molecular que realiza o transporte de lipídios (gorduras) em meios líquidos.

matriz estequiométrica: matriz esparsa que relaciona reações químicas (colunas) e compostos químicos (linhas) e contém os coeficientes estequiométricos para cada reagente ou produto em cada reação química.

MCMC (*Markov-Chain-Monte-Carlo*): técnica de amostragem de uma distribuição de probabilidade a partir de uma cadeia de Markov que apresenta a distribuição desejada no limite.

medida profilática: uma medida de precaução para evitar a proliferação de doenças, e.g., a vacinação.

metabolismo: o conjunto de reações químicas mais diretamente envolvido na manutenção do estado vivo de células e organismos.

metabolismo de sobrevivência: metabolismo do organismo em situações de estresse fisiológico, como exposição a antibióticos ou depleção de nutrientes.

metabólito: substância usada em um processo metabólico ou produzida durante o metabolismo.

metabolômica: estudo do *metaboloma*, i.e., do conjunto de todos os metabólitos presentes em um organismo ou célula.

micobactina: um sideroféro presente em micobactérias que transporta íons de ferro do ambiente extracelular para o citoplasma.

mRNA: ácido ribonucléico mensageiro; biomolécula sintetizada a partir do DNA a qual contém a 'mensagem' necessária para síntese de uma determinada proteína.

***operon*:** área do DNA contendo um conjunto de genes que são transcritos para uma única molécula de mRNA.

ortologia: similaridades genéticas em diferentes espécies derivadas de um mesmo ancestral.

parede celular: membrana rígida composta de polissacarídeos a qual envolve o plasma de diversos tipos de células como plantas, fungos e bactérias.

patógeno: microorganismo capaz de causar uma doença no organismo hospedeiro.

peptídeo: biomolécula formada pelo encadeamento de dois ou mais amino-ácidos.

peptidoglicano: polímero composto de açúcares e amino-ácidos presente na parede celular de diversas células.

pirimidina: composto orgânico presente na constituição do DNA.

polissacarídeo: moléculas de carbohidratos compostas principalmente de longas cadeias de açúcares simples.

procariótica: célula sem um núcleo definido.

produto natural: composto químico produzido por um organismo vivo. É encontrado na natureza ou pode ser produzido por síntese total e pode ser usado com propósitos farmacológicos ou biológicos.

proteína: macromoléculas responsáveis por diversas funções celulares como catálise, estrutura, replicação, etc.

proteômica: o estudo do proteoma, i.e., proteínas, suas funções, estruturas e interações.

purina: composto orgânico presente na constituição do DNA.

reação bioquímica: reação química que ocorre nos organismos vivos.

rede metabólica *in silico*: representação computacional (i.e. na forma de estruturas de dados) do metabolismo (parcial ou total) de uma célula, e que pode ser processada digitalmente por um computador.

reducionismo: técnica de estudo em que sistemas complexos são reduzidos a componentes mais simples para facilitar a sua análise e compreensão.

***regulon*:** área do DNA contendo um conjunto de genes que são regulados (inibidos ou ativados) de forma conjunta por uma mesma proteína. Em geral, o termo é aplicado a organismos procariontes.

respiração celular: processo celular metabólico de conversão da energia bioquímica presente nos nutrientes absorvidos pela célula.

sideroféro: composto orgânico responsável pela captação de ferro.

sistema de transporte ABC: conjunto de proteínas responsável pelo transporte de compostos e moléculas através da membrana celular em organismos procariontes.

teste-*t*: teste de hipótese estatístico que permite determinar se dois conjuntos de dados diferem de forma significativa. Neste procedimento a estatística de teste segue uma distribuição $\boldsymbol{t}$ de Student.

teste Welch: teste de hipótese estatístico similar ao teste-$\boldsymbol{t}$, porém apresentando maior confiabilidade no caso de amostras com tamanhos e variâncias populacionais diferentes.

transcrição: o primeiro passo no processo de duplicação que ocorre dentro do núcleo da célula. Aqui o DNA é dividido para produzir uma cópia (i.e. uma transcrição) do segmento desejado.

tradução: o segundo passo no processo de criação de proteínas. Ocorre fora do núcleo da célula. Neste passo, o mRNA (RNA mensageiro que foi enviado para fora do núcleo) é utilizado para sintetizar (i.e. traduzir) uma proteína.

valor-*p*: em teste de hipóteses, é a probabilidade de que seja obtido um valor mais extremo que o valor obtido quando a hipótese nula é verdadeira. Por exemplo, em testes de diferença de médias, um valor-p próximo de zero indica maior probabilidade de rejeitar a hipótese nula de que as médias não diferem significativamente.

ANEXO A – Listagem de Caminhos Metabólicos

Na tabela a seguir estão listados os caminhos metabólicos do *Mycobacterium tuberculosis* catalogados e disponíveis na base de dados KEGG(Kanehisa Labs, 2015)

Tabela 10 – Caminhos metabólicos para o organismo *Mycobacterium tuberculosis* catalogados na base de dados *Kyoto Encyclopaedia of Genes and Genomes* (KEGG).

mtu01110	Biossíntese de metabólitos secundários
mtu00010	Glicólise e gluconeogênesis
mtu00020	Ciclo do ácido tricarboxílico (ciclo TCA)
mtu00030	Via das Pentoses-fosfato
mtu00040	Interconversões Pentose e Glicuronato
mtu00051	Metabolismo de Frutose e Manose
mtu00052	Metabolismo de Galactose
mtu00053	Metabolismo de Ascorbato e Aldarato
mtu00500	Metabolismo de Sacarose e Amido
mtu00520	Metabolismo de Amino Açúcar e Açúcar de Nucleotídeo
mtu00620	Metabolismo de Piruvato
mtu00630	Metabolismo de Glioxilato e Dicarboxilato
mtu00640	Metabolismo de Propanoato
mtu00650	Metabolismo de Butanoato
mtu00660	Metabolismo de Ácido dibásico derivação 5C
mtu00562	Metabolismo de Fosfato de Inositol
mtu00190	Fosforilação Oxidativa
mtu00680	Metabolismo de Metano
mtu00910	Metabolismo de Nitrogênio
mtu00920	Metabolismo de Enxofre
mtu00061	Biossíntese de Ácidos Graxos
mtu00071	Metabolismo de Ácidos Graxos
mtu00072	Síntese e degradação de corpos derivados de cetonas
mtu00100	Biossíntese de Esteróides
mtu00561	Metabolismo de Glicerolipídeos
mtu00564	Metabolismo de Glicerofosfolipídeos
mtu00565	Metabolismo de Substratos Alcoóis-Graxos
mtu00592	Metabolismo do Ácido $\boldsymbol{\alpha}$-Linolênico
mtu01040	Biossíntese de Ácidos Graxos Não-saturados
mtu00230	Metabolismo de Purina
mtu00240	Metabolismo de Pirimidina
mtu00250	Metabolismo de Alanina, Ácido Aspártico e Glutâmico
mtu00260	Metabolismo de Glicina, serina e Treonina
mtu00270	Metabolismo de Cisteína e Metionina
mtu00280	Degradação de Valina, Leucina e Isoleucina
mtu00290	Biossíntese de Valina, Leucina e Isoleucina
mtu00300	Biossíntese de Lisina

mtu00310	Degradação de Lisina
mtu00330	Metabolismo de Arginina e Prolina
mtu00340	Metabolismo de Histidina
mtu00350	Metabolismo de Tirosina
mtu00360	Metabolismo de Fenilalanina
mtu00380	Metabolismo de Triptofano
mtu00400	Biossíntese de Fenilalanina, Tirosina e Triptofano
mtu00410	Metabolismo de β-Alanina
mtu00430	Metabolismo de Taurina e Hipotaurina
mtu00450	Metabolismo de Compostos Selênicos
mtu00460	Metabolismo de cianoamino-ácido
mtu00471	Metabolismo de D-Glutamina e D-glutamato
mtu00472	Metabolismo de D-Arginina e D-ornitina
mtu00473	Metabolismo de D-Alanina
mtu00480	Metabolismo de Glutationa
mtu00540	Biossíntese de Lipopolissacarídeos
mtu00550	Biossíntese de Peptidoglicano
mtu00730	Metabolismo de Tiamina
mtu00740	Metabolismo de Riboflavina
mtu00750	Metabolismo de Vitamina B6
mtu00760	Metabolismo de Ácido Nicotínico e Nicotinamida
mtu00770	Biossíntese de Pantotenato e Coenzima-A
mtu00780	Metabolismo de Biotina
mtu00785	Metabolismo de Ácido Lipóico
mtu00790	Biossíntese de Folato
mtu00670	Via do Folato e Carbono-1
mtu00860	Metabolismo de Porfirina e Clorofila
mtu00130	Biossíntese de Ubiquinona e Terpenóides-Quinonas
mtu00900	Biossíntese da Estrutura de Terpenóides
mtu00906	Biossíntese de Carotenóides
mtu00903	Degradação do Limoneno e do Pineno
mtu00281	Degradação do Geraniol
mtu00523	Biossíntese da Unidade de Açúcar de Policetídeos
mtu01053	Biossíntese de Sideróforos e Peptídeos Não-ribossômicos
mtu00311	Biossíntese de Penicilina e Cefalosporina
mtu00521	Biossíntese de Estreptomicina
mtu00401	Biossíntese da Novobiocina
mtu00362	Degradação do Benzoato
mtu00627	Degradação do Aminobenzoato
mtu00364	Degradação do Fluorbenzoato
mtu00625	Degradação do Cloroalcano e Cloroalceno
mtu00361	Degradação do Clorociclohexano e Clorobenzeno
mtu00623	Degradação do Tolueno

mtu00622	Degradação do Xileno
mtu00633	Degradação do Nitrotolueno
mtu00642	Degradação do Etilbenzeno
mtu00643	Degradação do Estireno
mtu00930	Degradação do Caprolactama
mtu00363	Degradação do Bisfenol
mtu00621	Degradação da Dioxina
mtu00626	Degradação do Naftaleno
mtu00624	Degradação de Hidrocarbonetos Policíclicos Aromáticos
mtu00984	Degradação de Esteróides
mtu01210	Metabolismo do Ácido 2-Oxocarboxílico
mtu01220	Degradação de Compostos Aromáticos
mtu01230	Biossíntese de Amino-ácidos
mtu03020	RNA polimerase
mtu03010	Ribossomos
mtu00970	Biossíntese de tRNA-Aminoacil
mtu03060	Exportação de Proteínas
mtu04122	Sistema de Trocas com Enxofre
mtu03050	Proteasoma
mtu03018	Degradação de RNA
mtu03030	Replicação de DNA
mtu03410	Reparação de Excisão de Bases
mtu03420	Reparação de Excisão de Nucleotídeos
mtu03430	Reparação de bases Mal-pareadas
mtu02010	Transportadores ABC
mtu03070	Sistema de Secreção Bacterial
mtu02020	Sistema de Transporte Dicomponente

Esta tabela mostra a lista de caminhos metabólicos para a bactéria *Mycobacterium tuberculosis* catalogados na base de dados KEGG. Os caminhos são de biossíntese e degradação de compostos dentro do ambiente celular. Alguns caminhos como o mtu03018 e mtu03420 incluem funções de reparação de DNA e controle transcripcional, e ainda outros contém reações e enzimas reponsáveis pelos sistemas de transporte através da membrana como mtu02010 e mtu03070. A primeira coluna indica o código do caminho metabólico na base de dados e a segunda coluna uma descrição do objetivo metabólico final do caminho.

ANEXO B – Tratamentos Boshoff

Na tabela a seguir estão listados os tratamentos e compostos antibióticos estudados em (BOSHOFF et al., 2004) com o *Mycobacterium tuberculosis*. São 75 diferentes tratamentos avaliados em diversas concentrações, condições de crescimento e tempos de exposição, na forma de um conjunto de 437 micromatrizes *dual-chip*.

Tabela 11 – Tratamentos e compostos antibióticos avaliados no conjunto de dados de micromatrizes de DNA em (BOSHOFF et al., 2004).

ID	Sigla	Descrição do tratamento	Classe
01	#109	0,1-10 μg/mL Diamine analog 109	1
02	#111891	1-5 μg/mL synthetic pyridoacridine analog	2
03	#111895	1-5 μg/mL ascididemin	2
04	#121940	1-5 μg/mL synthetic pyrodoacridine analog	2
05	#124196	1-5 μg/mL synthetic pyrudoacridine analog	2
06	#241	0,1-10 μg/mL Diamine analog 241	1
07	#592	0,1-10 μg/mL Diamine analog 59	1
08	5CL-PZA	40-80 μg/mL 5-Chloropyrazinamide	3
09	Amikacin	5-10 μg/mL Amikacin	4
10	Amp	0,2 μg/mL Ampicillin	1
11	ARP4	1-4 μg/mL antitubercular compound ARP4	***
12	Antimycin	12,5-50 μg/mL Antimycin A	***
13	Asc Nat Prod	4-9 μg/mL extract of *Eudistoma amplum*	2
14	BZA	0,12 mg/mL benzamide	2
15	Cap	5-10 μg/mL capreomycin	4
16	CCCP	10-50 μM c. cyan. 3-chlorophenylhydrazone	5
17	Cephalexin	20-100 μg/mL cephalexin	1
18	Cerulenin	0,32-5 μg/mL cerulenin	1
19	Clofazimine	10-13 μg/mL clofazimine	5
20	Clotrimazole	10-24 μg/mL clotrimazole (azole drug)	5
21	CPZ	10-50 μg/mL chlorpromazine	5
22	DCCD	20-100 μM dicyclohexylcarbodiimide	5
23	Deferoxamine	150-250 μM deferoxamine mesylate	2
24	DIPED	5-100 μg/mL diisopropylethylenediamine	1
25	Dipyridyl	100-200 μM 2,2'-Bipyridyl	2
26	DNP	0,5-2,5 mM 2,4-dinitrophenol	5

ID	Sigla	Descrição do tratamento	Classe
27	DTNB	1-2 mM 5,5'-dithiobis (2-nitrobenzoic acid)	***
28	DTT	1-2 mM 1,4-dithio-DL-threitol	5
29	Econazole	10-24 μg/mL econazole (azole drug)	5
30	EMB	10-20 μg/mL ethambutol	1
31	Ethionamide	12-40 μg/mL ethionamide	1
32	GSNO	0,02-0,1 mM S-Nitrosoglutathione	6
33	GSNO/CPZ	0,1 mM GSNO + 20-25 μg/mL CPZ	5
34	GSNO/CFZ	0,1 mM GSNO + 5 μg/mL CFZ	5
35	GSNO/KCN	0,1 mM GSNO + 5 μg/mL KCN	5
36	GSNO/Med	0,1 mM GSNO + 6-10 μg/mL menadione	5
37	H_2O_2	4 mM hydrogen peroxide	7
38	INH	0,2-0,4 μg/mL isoniazid	1
39	KCN	5-20 μg/mL potassium cyanide	5
40	Levo	10 μg/mL levofloxacin	7
41	Menadione	6-10 μg/mL menadione	5
42	Mercaptoethanol	2 mM 2-Mercaptoethanol	***
43	Methoxatin	10-20 μg/mL methoxatin	***
44	MinMedSuc	minimal medium with succinate	***
45	MTM	0,1-0,2 μg/mL mitomycin C	7
46	NAM	0,12-1,2 mg/mL Nicotinamide	3
47	NaN_3	0,2-2 mM sodium azide	5
48	Nigericin	50 μg nigericin	5
49	Novobiocin	10 μg/mL novobiocin	7
50	NRP-1	non-replicating persistence stage 1	6
51	Oflox	5-10 μg/mL ofloxacin	7
52	PA-1	10-50 μg/mL antiTB PA-1	***
53	PA-21	10-50 μg/mL antiTB PA-21	***
54	PA-824	0,2-2 μg/mL antiTB PA-824	***
55	MinMedPal	minimal medium with palmitate	***
56	pH4.8	pH 4,8 acidic medium	9
57	pH5.2	pH 5,2 acidic medium	9
58	pH5.6	pH 5,6 acidic medium	9
59	Procept 6776	20-30 μg/mL antiTB Procept 6776	***
60	Procept 6778	5-40 μg/mL antiTB Procept 6778	***

ID	Sigla	Descrição do tratamento	Classe
61	PZA	0,12-1,2 mg/mL pyrazinamide	3
62	Rif	0,2-5 μg/mL rifampicin	8
63	Rifp	0,1-0,5 μg/mL rifapentine	8
64	Rox	10-50 μg/mL roxithromycin	4
65	STREP	2-5 μg/mL streptomycin	4
66	starvation	Starvation in PBS T80 medium	10
67	MinMedSuc	minimal medium succinate vs. glucose	10
68	Tet	5-10 μg/mL tetracycline	4
69	TLM	0,1-0,2 mg/mL thiolactomycin	1
70	TRC	10-150 μg/mL triclosan	5
71	TRZ	10-25 μg/mL thioridazine	5
72	UV	UV irradiation	7
73	Valinomycin	0,5-10 μM valinomycin	5
74	Verapamil	50 μM verapamil	***
75	$ZnSO_4$	2 mM $ZnSO_4$	5

Esta tabela mostra a lista de tratamentos e compostos analisados no conjunto Boshoff. Primeira coluna indica a sigla utilizada no artigo original; segunda coluna apresenta uma descrição do tratamento e condições experimentais utilizadas. A terceira coluna indica a classe de ação bactericida indica em (BOSHOFF et al., 2004) de acordo com a seguinte numeração: (1) inibição da síntese da parede celular; (2) piridoacridonas e sequestrantes de ferro; (3) amidas aromáticas hidrolizadas no ambiente intracelular; (4) inibição de síntese protéica; (5) inibição da respiração por agentes diferentes de NO; (6) crescimento sob condições associadas com a expressão do *regulon* DoS (7) agentes causadores de danos à integridade ou topologia do DNA (8) inibidores transcripcionais; (9) meio ácido; (10) falta de nutrientes. Os tratamentos indicados por *** não possuem classificação *a priori* apresentada em (BOSHOFF et al., 2004).

ANEXO C – Biossíntese de Metionina

A Figura 38 está disponível para visualização e *download* a partir da base de dados KEGG (Kanehisa Labs, 2015), e mostra enzimas e metabólitos envolvidos na biossíntese (i.e. produção) dos amino-ácidos cisteína e metionina no organismo *Mycobacterium tuberculosis*.

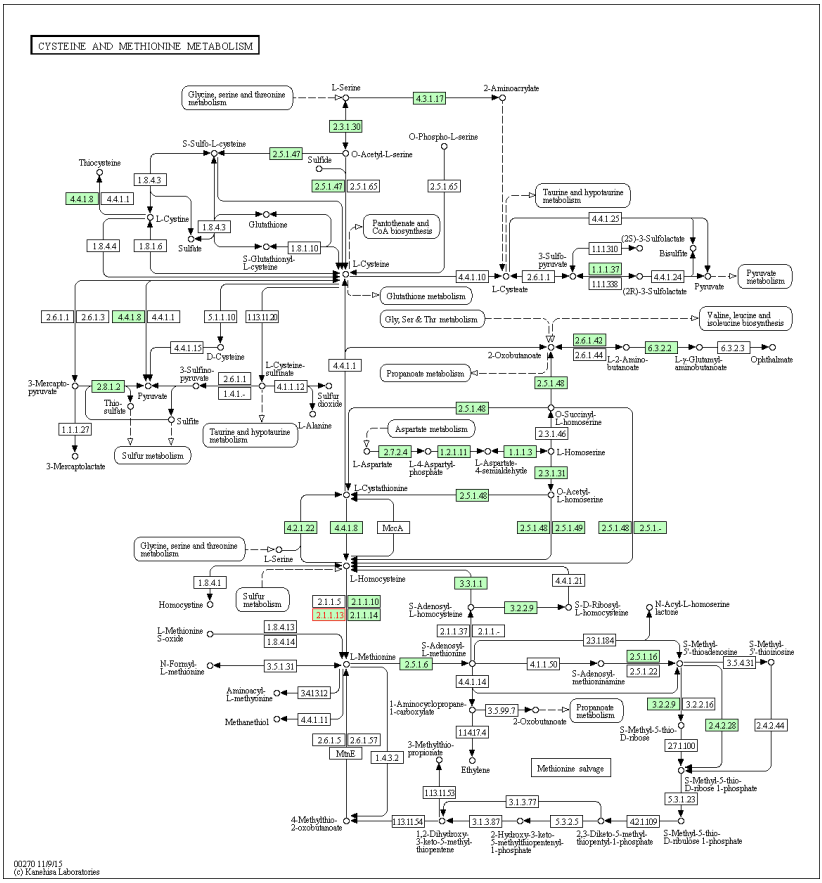

Figura 38 – Estrutura conhecida do caminho metabólico (*mtu00270*) para biossíntese dos amino-ácidos metionina e cisteína no organismo *Mycobacterium tuberculosis*. Retângulos com cantos arredondados representam associações e interações com outros caminho metabólicos, retângulos representam enzimas e reação metabólica associada, círculos representam metabólitos (reagentes e produtos das reações), setas indicam direcionalidade da reação. Observa-se que algumas enzimas catalisam apenas em uma direção. *Fonte: (Kanehisa Labs, 2015).*