UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM
COMPUTAÇÃO APLICADA

Redes Neurais Recorrentes para Inferência de Redes de Interação Gênica utilizando Cadeias de Markov

por

ÍGOR LORENZATO ALMEIDA

Dissertação submetida a avaliação como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada

Orientador: Prof Dr. Adelmo Luis Cechin

São Leopoldo, Janeiro de 2007

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Almeida, Ígor Lorenzato

Redes Neurais Recorrentes para Inferência de Redes de Interação Gênica utilizando Cadeias de Markov / por Ígor Lorenzato Almeida. — São Leopoldo: Ciências Exatas e Tecnologicas da UNISINOS, 2007.

108 f.: il.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos. Ciências Exatas e Tecnológicas Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, São Leopoldo, BR–RS, 2007. Orientador: Cechin, Adelmo Luis.

1. RNA. 2. Microarranjos. I. Cechin, Adelmo Luis. II. Título.

Sumário

Lista de Abreviaturas

CDC Controle de Divisão Celular

CTA Ciclo do Ácido Carboxílico

DBN Redes Bayesianas Dinâmicas

HMM Modelo Oculto de Markov

PCA Análise de Componentes Principais

PCR *Polymerase Chain Reaction*

RAC *Rank and Combinatory Analysis*

RNA Rede Neural Artificial

RNN *Recurrent Neural Network*

RNR Rede Neural Recorrente

SMD *Stanford Microarray Database*

SOM Mapas Auto-Organizáveis

SVD *Singular Value Decomposition*

SVM Máquinas de Suporte Vetorial

Lista de Figuras

Agradecimentos

Ao meu orientador, Prof. Dr. Adelmo Luis Cechin, não apenas pelo conhecimento passado e pela orientação concedida, mas principalmente pela forma de fazê-lo, respeitando sempre minha opinião e estando sempre disponível para boas discussões durante as noites de sexta-feira.

Aos professores, Dr. Luiz Paulo Luna de Oliveira e Dr. Arthur Tórgor Gómez, pelas importantes contribuições dadas no decorrer deste trabalho.

Aos meus pais, por estarem sempre presentes, confiando em mim e apoiando incondicionalmente minhas decisões.

Aos meus amigos que contribuiram, cada um à sua forma, com a realização deste trabalho.

À Hewlett-Packard Brasil P&D pela bolsa de estudos concedida, sem a qual a realização deste trabalho não seria possível.

E principalmente à Denise Regina Pechmann, pelo carinho, companheirismo e apoio.

Resumo

Microarranjos têm sido fortemente usados para monitorar, simultaneamente, o padrão de expressão de milhares de genes. Assim, uma grande quantidade de dados tem sido gerada e o desafio atual é descobrir como extrair informações úteis destes conjuntos de dados.

Dados de Microarranjos são fortemente especializados, envolvendo diversas variáveis de forma não linear e temporal, necessitando de modelos recorrentes não lineares, os quais são complexos para formular e analisar. Este trabalho propõe a utilização de Redes Neurais Recorrentes (RNR) como modelo para os dados devido às suas habilidades de aprendizado de sistemas não-lineares e complexos.

Uma vez obtido um modelo para os dados utilizando uma RNR, é possível extrair regras que representam as características aprendidas. Analisando as regras em conjunto com a base de dados, propõe-se a representação do conhecimento utilizando Cadeias de Markov. Tais Cadeias são facilmente visualizadas, na forma de grafos de estados, apresentando as interações entre os níveis de expressão gênica, bem como seus padrões temporais.

Assim, é proposta uma nova abordagem para análise de dados de Microarranjos, através da extração de Cadeias de Markov, a partir de RNRs. Dois aspectos importantes são analisados: a evolução no tempo da expressão gênica e sua influência mútua na forma de redes regulatórias. Com isto pretende-se fornecer aos especialistas, de uma forma simples, indícios sobre possíveis relações de causa e conseqüência presentes nos dados, bem como sobre os padrões temporais existentes, tudo em uma forma de fácil compreensão.

TITLE: "RECURRENT NEURAL NETS FOR NETWORKS INFERENCE OF GENE INTERACTIONS USING MARKOV CHAINS "

Abstract

Array technologies have made it straightforward to simultaneously monitor the expression pattern of thousands of genes. Thus, a lot of data is being generated and the challenge now is to discover how to extract useful information from these data sets.

Microarray data is highly specialized. It involves several variables in a nonlinear and temporal way, demanding nonlinear recurrent free models, which are complex to formulate and to analyse. So, this work proposes the use of Recurrent Neural Networks (RNN) for data modeling, due to their learning hability of nonlinear and complex systems.

Once a model is obtained with a RNN for the data, it is possible to extract rules to represent the knowledge acquired by them. From rule analisys, this work proposes the representation of the knowledge by Markov Chains model, which is easily visualized in the form of a graph of states, which show the interactions among the gene expression levels and their changes in time

In this work, we propose a new approach to microarray data analysis, by extracting a Markov Chain from trained RNNs. Two aspects are of interest for the researcher: the time evolution of the genic expression and their mutual influence in the form of regulatory networks. This work aims at providing some relevant information about possible cause and consequences relations among genes, in a simple way, to domain experts.

Capítulo 1

Introdução

Com o crescente avanço dos projetos de seqüenciamento genômico, um grande volume de dados vem sendo gerado, entre eles os dados de Microarranjos. Mediante isto, a necessidade de técnicas computacionais eficazes para análise desses dados tem se tornado cada vez maior.

Os Microarranjos foram desenvolvidos na década de 90 e visam analisar a expressão gênica de milhares de genes em conjunto. Com eles é possível o estudo de padrões de expressão gênica, os quais fundamentam a fisiologia celular, analisando a ativação (ou inativação) dos genes em um determinado meio [Alberts et al. 2004, Causton et al. 2003, Kohane et al. 2003]. Essa informação sobre o comportamento temporal destes genes, em um mesmo ambiente, permite fazer inferências importantes sobre sua interação. Tais inferências podem descrever como todo o conjunto de genes reage ao ambiente submetido. Isto pode ser utilizado, por exemplo, para descobrir novas drogas para combater doenças e também diminuir efeitos colaterais de drogas já existentes.

Extrair tais relações de um conjunto de dados, como os Microarranjos não é uma tarefa simples. Estes dados são séries temporais com muitas variáveis e poucas amostras temporais. Muito tem sido feito na área utilizando técnicas como Redes Bayesianas ou, até mesmo, inferência matemática [Noman and Iba 2005, Friedman 2004, Qian et al. 2003, Zou and Conzen 2005]. No entanto, ainda não foi desenvolvida uma metodologia para, somente com dados de Microarranjos, obter a relação de interações entre os genes, na forma de uma Rede Biológica, representando o comportamento dos genes durante todo o período amostrado.

Desta forma, visando sanar a carência de técnicas para este tipo de análise, este trabalho visa propor uma nova metodologia para identificar as interações existentes entre os genes a partir de dados de Microarranjos. Para este objetivo é aplicada a metodologia de extração de conhecimento a partir de Redes Neurais Recorrentes (RNR), apresentada em [Pechmann and Cechin 2005, Pechmann and Cechin 2004, Cechin 1997], utilizando os dados de Microarranjos para o treinamento das redes e, representado o conhecimento extraído na forma de Cadeias de Markov.

Para atingir este objetivo, o trabalho se divide em diversas etapas, iniciando pela determinação da base de dados a ser utilizada. Após a determinação da base de dados que apresenta as relações de interesse para o trabalho, escolhida entre as

inúmeras disponíveis na internet, é necessário processar os dados originais para construir a base de treinamento da RNR. Microarranjos possuem milhares de variáveis, e as RNRs não suportam tal dimensão de dados, muitas vezes não conseguindo modelar tamanha quantidade de informação, sendo necessário reduzir o número de variáveis a ser utilizada. Treinar a RNR é a etapa seguinte, na qual estipulam-se os parâmetros topológicos e de treinamento para a rede modelar os dados de forma satisfatória. Após o treinamento da rede, são extraídas informações sobre seu aprendizado na forma de uma cadeia de Markov, a qual representa os diversos estados de interação presente no Microarranjos, cada um destes estados descrevendo as interações entre os genes na forma de uma rede regulatória.

1.1 Objetivos

Os dados gerados por Microarranjos são séries temporais, cada série correspondendo ao nível de expressão de um gene, ao longo do tempo, em um certo meio. Em um único experimento pode ser analisado a expressão de milhares de genes, no entanto a interpretação dos dados resultantes se torna uma tarefa complexa, principalmente pelo volume de informações presente.

1.1.1 Objetivo Geral

O objetivo geral deste trabalho consiste em desenvolver uma metolodogia de Extração de Conhecimento para extrair as possíveis interações existentes entre os genes da análise a partir de dados brutos de Microarranjos. Estas interações serão apresentadas na forma de uma Cadeia de Markov e Sistemas Lineares representando Redes Regulatórias.

Desta forma, pretende-se prover aos especialistas indícios sobre as possíveis relações de causa e conseqüência existentes entre os genes, além de identificar padrões temporais, os genes mais importantes, bem como, as fases do processo biológico analisado, tudo isto apresentado em uma forma de fácil compreensão.

1.1.2 Objetivos Específicos

Como objetivos específicos deste trabalho, podem ser citados:

- uso de RNRs como forma de modelo não-linear pra dados de Microarranjos;
- através de Cadeias de Markov, representar as diferentes interações temporais presentes nos dados;
- obter um sistema linear que defina as interações entre os genes possibilitando a definição de uma Rede Regulatória.

1.2 Relevância do Estudo

Através da representação do comportamento temporal da expressão gênica por meio de Cadeias de Markov, busca-se encontrar as relações e influências existentes entre os genes ou conjunto deles. Este tipo de conhecimento é valioso para os biólogos, pois possibilita uma melhor compreensão sobre o organismo em análise. Primeiramente, através do entendimento da rede de ativação gênica é possível predizer a reação do organismo (genes ativados) sob influência de drogas. Em segundo lugar, a relação temporal na forma de Cadeias de Markov permite ao cientista entender e predizer em que condições o organismo muda sua rede de níveis de expressão gênica, e conseqüentemente como ele se adapta às condições adversas. Extrair tais informações, sem o uso de técnicas computacionais, torna-se inviável, mediante o volume de dados em questão.

1.3 Apresentação do Texto

O texto deste trabalho está divido como segue. O Capítulo 2 apresenta conceitos básicos de Biologia Molecular referentes aos dados utilizados neste trabalho. Apresenta uma breve introdução sobre expressão gênica e sua regulação, a qual pode ser tanto positiva, quanto negativa. Neste capítulo é apresentada a técnica de medida do nível de expressão gênica com o uso de Microarranjos, define os Microarranjos, apresenta o tipo de dado produzido além de citar algumas análises clássicas realizadas sobre eles.

Os conceitos sobre as técnicas computacionais utilizadas neste trabalho são abordados no capítulo 3. São apresentados os fundamentos das metodologias de Extração de Conhecimento e Mineração de Dados. Após uma breve explanação sobre RNAs e SOM, finalizando o Capítulo com os fundamentos da Cadeias de Markov.

No Capítulo 4, consta uma revisão bibliográfica citando artigos relacionados ao problema em questão, seguido, no Capítulo 5, pela descrição da metodologia proposta, explicando os fatores levados em consideração para a escolha da base de dados, as técnicas utilizadas para o pré-tratamento e seus objetivos. Explica, em detalhes, o treinamento e avaliação de desempenho da RNR utilizada para modelar os dados. Neste Capítulo é explicada a extração do conhecimento e a representação deste por uma Cadeia de Markov.

A descrição dos experimentos realizados, os resultados obtidos e uma breve discussão a respeito de seu significado e validade, constam no Capítulo 6, seguido, no Capítulo 7, pelas conclusões e possibilidades de trabalhos futuros. Ao final constam as referências bibliográficas utilizadas neste trabalho.

Capítulo 2

Conceitos Básicos sobre Biologia Molecular

A Biologia Molecular, como qualquer outra ciência moderna, depende de instrumentos para dissecar a arquitetura e a operação de sistemas inacessíveis aos sentidos humanos.

Além do instrumental físico e químico para separar, quantificar e analisar o material biológico, os biólogos moleculares também se valem de recursos computacionais cada dia mais sofisticados.

Neste capítulo será apresentada uma compilação dos principais conceitos sobre a Biologia Molecular necessários à compreensão deste estudo.

2.1 Expressão Gênica

Nesta seção serão abordados alguns conceitos sobre expressão gênica. Tais conceitos são aceitos pela comunidade científica e estão baseados nos trabalhos de [Lehninger et al. 1995, Voet et al. 2000, Alberts et al. 2004].

A expressão da informação genética contida num segmento de DNA sempre envolve a geração de uma molécula de RNA. O RNA é a única molécula conhecida a ter funções tanto informacionais quanto catalíticas, levando a inúmeras especulações de que ela possa ter sido o intermediário químico essencial para o desenvolvimento da vida no planeta.

Com exceção de certos vírus, todas as moléculas de RNA são derivadas da informação permanente armazenada no DNA. Assim, o processo de transcrição, realizado pela enzima RNA polimerase, é a conversão da informação genética de um segmento de DNA em uma fita de RNA com uma seqüência de bases complementar a uma das fitas do DNA. Entre as espécies de RNA produzidas, três são consideradas principais: RNA Mensageiro (mRNA), RNA de transferência (tRNA) e o RNA Ribossômico (rRNA).

A transcrição é dita um processo seletivo, onde apenas genes ou grupos de genes particulares são transcritos durante um certo tempo. Desta forma, a transcrição do DNA pode ser regulada de forma que apenas a informação genética necessária

para a célula em um momento particular seja transcrita.

2.1.1 Regulação da Expressão Gênica

Dos milhares de genes presentes em um organismo, apenas uma fração deles é expressa em um mesmo instante de tempo. Alguns produtos gênicos (moléculas de RNA ou cadeias polipeptídicas que apresentam uma função específica na célula), oriundos da expressão de um gene, têm funções que demandam sua presença em quantidades muito grandes, outros nem tanto. As necessidades de um certo produto gênico podem também se alterar com o tempo. Desta forma, a regulação da expressão gênica é um componente crítico na regulação do metabolismo celular e na manutenção das diferenças estruturais e funcionais que existem nas células durante o desenvolvimento. A regulação gênica também é essencial para a célula minimizar o gasto de energia disponível.

Regular a concentração de uma proteína envolve um equilíbrio de diversos processos, onde há pelo menos seis pontos potenciais onde a quantidade de proteína pode ser regulada, conforme citados abaixo e apresentados na Figura 2.1 (extraída de Lehninger [Lehninger et al. 1995]).

- síntese do transcrito primário do RNA a partir do DNA;
- processamento pós-transcrição do mRNA;
- degradação do mRNA;
- tradução ou síntese protéica, ponto em que ocorre a conversão do mRNA em proteína;
- modificação pós-tradução e
- degradação protéica.

A concentração de uma dada proteína pode ser regulada em qualquer ou até mesmo em todos estes processos.

O grau e o tipo de regulação refletem a função da proteína codificada pelo gene. Alguns produtos são requeridos todo o tempo, logo seus genes sofrem uma regulação a fim de serem expressos de maneira constante em praticamente todas as células da espécie ou organismo. Tal expressão é chamada de **constitutiva**. Há produtos gênicos que possuem concentrações variantes a sinais moleculares, podendo esta aumentar ou diminuir frente a alguma condição. Produtos que aumentam sua concentração sofrem um processo de **indução** (aumento da expressão gênica), já os que diminuem sua concentração, sofrem **repressão** (diminuição da expressão gênica).

A transcrição é mediada e regulada por interações proteína-DNA, cujo componente principal é a RNA polimerase, a qual se liga ao DNA iniciando a transcrição em sítios específicos, chamados de promotores. A iniciação da transcrição é regulada por proteínas, que se ligam nos promotores ou próximo deles. Há, pelo menos, três tipos de proteínas envolvidas nesta regulação:

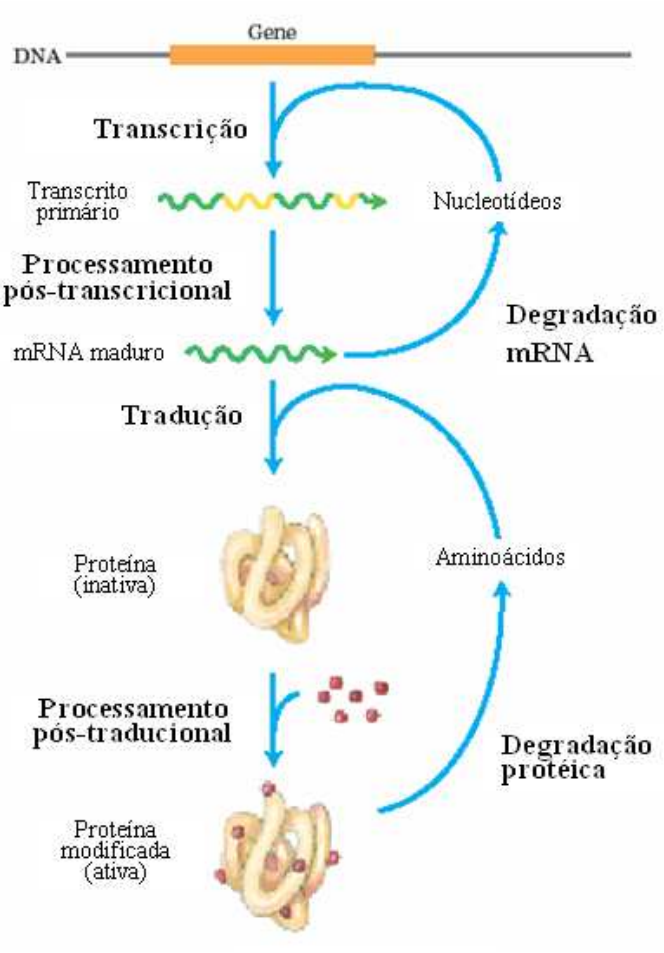


FIGURA 2.1 – *Seis processos que são pontos potencias para regulação.*

- **Fatores de especificidade:** alteram a especificidade da RNA polimerase para um certo promotor, ou conjunto deles.
- **Repressores:** ligam-se a um promotor bloqueando o acesso da RNA polimerase a ele.
- **Ativadores:** ligam-se próximos de um promotor, aumentando a interação RNA-promotor.

A regulação gênica apresenta dois tipos: regulação negativa e regulação positiva. Estes tipos de regulação serão abordados a seguir.

2.1.2 Regulação Negativa

Este tipo de regulação ocorre quando os repressores ligam-se a sítios específicos do DNA, os operadores. Os sítios operadores estão, geralmente, próximos e, freqüentemente, se sobrepõem ao promotor, de tal forma que a ligação da RNA polimerase,

ou a sua movimentação ao longo do DNA após a ligação, seja bloqueada, sempre que o repressor estiver presente.

A ligação do repressor é regulada por um sinal molecular, usualmente uma pequena molécula específica que se liga e induz uma alteração conformacional no repressor. A interação entre o repressor e a molécula sinal pode levar tanto a um aumento, quanto a uma diminuição na transcrição. Há casos em que o sinal bloqueia a ação do repressor, permitindo o início da transcrição, mas há sinais que agem nos repressores inativos, forçando-os a se ligarem ao operador.

2.1.3 Regulação Positiva

A regulação positiva ocorre quando é mediada por um ativador. Ativadores são um contraponto molecular aos repressores. Os ativadores ligam-se a sítios adjacentes a um promotor e aumentam a ligação e a atividade da RNA polimerase naquele promotor.

Os sítios de ligação para os ativadores são freqüentemente encontrados adjacentes aos promotores, estes em geral, ligam-se fracamente à RNA polimerase. A transcrição nestes genes é freqüentemente negligenciável na ausência do ativador. Algumas vezes o ativador está ligado ao DNA e dissocia-se quando ocorre a ligação com molécula sinal, geralmente uma pequena molécula específica, ou outra proteína. Quando ligada ao DNA, a proteína ativadora facilita a ligação da RNA polimerase e aumenta a velocidade da iniciação da transcrição. Em outros casos, o ativador somente se liga ao DNA após a ligação a uma molécula sinal.

2.2 Microarranjos

Os Microarranjos de DNA, desenvolvidos nos anos 90, revolucionaram a maneira de analisar a expressão gênica, permitindo que os produtos de RNA de milhares de genes sejam monitorados de uma só vez. Examinando a expressão de tantos genes simultaneamente, pode-se identificar e estudar os padrões de expressão gênica que fundamentam a fisiologia celular [Alberts et al. 2004].

Entre as aplicações de Microarranjos, a indústria farmacêutica ocupa um grande espaço, estudando o comportamento dos organismos a fim de determinar sua reação a estímulos externos, tais como medicamentos, podendo assim desenvolver estratégias para reduzir os efeitos colaterais no tratamento de doenças.

A seguir serão comentados alguns aspectos dessa técnica: uma breve explicação a respeito de seu funcionamento, bem como do tipo de dados que produz, suas características e algumas técnicas para sua análise.

2.2.1 Definição de Microarranjos

Microarranjos de DNA são lâminas de microscópio crivadas com uma grande quantidade de fragmentos de DNA, cada um contendo uma seqüência de nucleotídeos, que serve como uma sonda para um gene específico. Os arranjos mais densos devem conter dezenas de milhares destes fragmentos em uma área menor do

que um selo, permitindo que milhares de reações de hibridização sejam realizadas em paralelo. Alguns microarranjos gerados a partir de fragmentos grandes de DNA, foram gerados por PCR[1] e, então, plotados em lâminas, por um robô. Outros contém oligonucleotídeos curtos que são sintetizados na superfície de uma pastilha de vidro com técnicas simulares àquelas utilizadas para gravar circuitos em *chips* de computador. Em cada caso existe uma identificação de cada sonda no *chip*. Desta maneira, quaisquer fragmentos de nucleotídeo que hibridizem com uma sonda no arranjo, podem ser identificados como produtos de um gene específico, simplesmente detectando a posição à qual ela se liga [Alberts et al. 2004, Kohane et al. 2003].

Para utilizar um microarranjo de DNA para monitorar a expressão gênica, o mRNA das células que estão sendo estudadas é primeiro extraído e convertido para cDNA, e então marcado com uma sonda fluorescente. Em seguida o Microarranjo é encubado com a amostra de cDNA marcada e se realiza a hibridização. Com a hibridização alguns segmentos de DNA ligam-se às posições do Microarranjo, e estas posições são identificadas por um microscópio automático de varredura a *laser*. A seguir as posições dos arranjos são então comparadas com a do gene específico, do qual a amostra de DNA foi plotada. A Figura 2.2 mostra este processo na forma de esquema.

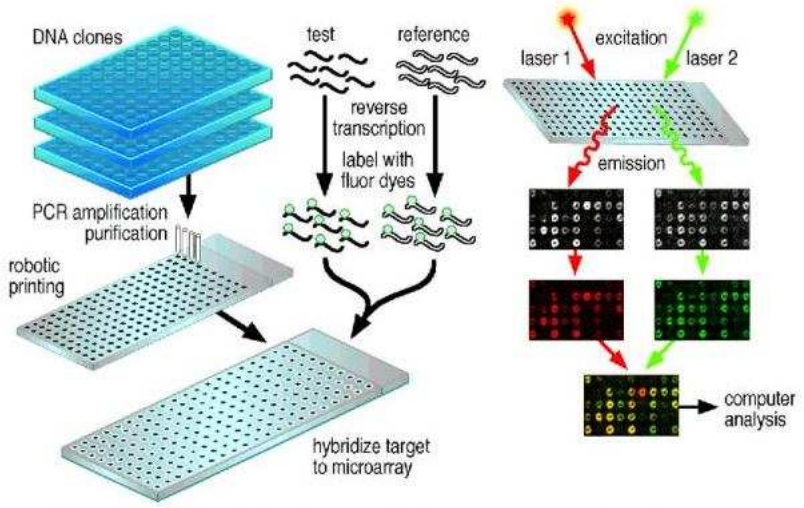


FIGURA 2.2 – *Esquema de um Microarranjo [Duggan et al. 1999].*

A aplicação mais popular dos microarranjos é comparar o nível de expressão em duas amostras diferentes: o mesmo tipo de célula sob duas condições [Causton et al. 2003]. Isto se baseia na identificação das amostras de mRNA extraídas com dois marcadores de cores diferentes: verde se a amostra é oriunda da condição 1 e vermelho se da condição 2, por exemplo. Assim, o microarranjo é submetido a um *laser*, respondendo com diferentes níveis de fluorescência de acordo com o estado no qual se encontra. Esta quantidade de fluorescência corresponde a

[1]PCR - *Polymerase Chain Reaction*, técnica utilizada para amplificar regiões do DNA cujas sequências são conhecidas

quantidade de ácido nucléico ligado a cada spot. Se a quantidade de ácido nucléico do experimento 1 estiver abundante, o spot será verde, já se for o do experimento 2, então será vermelho. Caso ambos sejam iguais, a resposta será amarela, e por fim, se não houver a presença de ácido nucléico, ele não apresentará fluorescência e resultará em um spot preto. A Figura 2.3 mostra um exemplo de um Microarranjo após ter sido submetido ao *laser*.

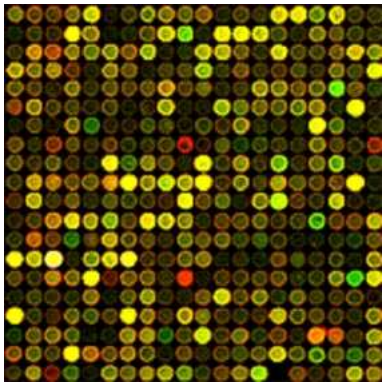


FIGURA 2.3 – *Região de um Microarranjo representando o nível de expressão de genes.*

Então, de acordo com o nível de fluorescência e cores de cada spot, o nível de expressão relativa dos genes pode ser estimado. Dessa maneira milhares de dados podem ser obtidos sobre a expressão gênica de um caso particular em um único experimento.

Após obter os níveis de fluorescência (quantização), começa a parte de análise dos dados gerados pelo experimento. Uma das características marcantes quando se trabalha com microarranjos é a grande quantidade de dados gerada. Frente a isso, um trabalho intenso de pré-processamento e normalização das informações é necessário, a fim de saber o que é relevante para ser analisado, principalmente, retirar os ruídos dos experimentos, e também deixar os dados em condição para comparação.

A forma final dos dados de microarranjos é a matriz de expressão gênica, da qual muitas informações podem ser extraídas, tais como aglomerados de genes com o mesmo comportamento, ou genes que se comportam diferentemente da maioria. Tanto um tipo de informação quando o outro pode ser importante, dependendo da análise realizada. Mas este é apenas um exemplo do que pode ser feito com estes dados. A análise destas matrizes é um desafio para muitos pesquisadores.

Entre os desafios encontrados nesta área de pesquisa está a determinação de uma metodologia para definir o tipo de informaçõe a extrair desses dados e, através de técnicas computacionais, determinar a relação existente entre os genes. Desta forma, os Microarranjos se tornam um interessante alvo de pesquisas sobre técnicas e métodos computacionais, sendo enfoque deste trabalho a determinação das relações existentes entre os genes.

2.2.2 Manipulação dos Dados

Os Microarranjos, devido a complexidade de seus dados, se tornam uma aplicação interessante para estudos sobre técnicas de Inteligência Artificial, envolvendo principamente técnicas de Mineração de Dados e Aprendizado de Máquina.

A seguir serão citadas algumas análises típicas realizadas sobre Microarranjos e também algumas técnicas empregadas para cada uma delas:

- *Clusterização:* O objetivo desta análise é agrupar dados semelhantes presentes no Microarranjo. Para isto diversas técnicas podem ser aplicadas, cada uma apresentando diferentes características em seus resultados. Entre as técnicas mais utilizadas estão:

  - *Clusterização Hierárquica:* esta técnica reordena todos os dados de forma a agrupar os semelhantes, fazendo com que a distância entre os dados seja proporcional a diferença entre eles [Eisen et al. 1998, Spellman et al. 1998].
  - *Vizinhos mais próximos:* a partir de valores iniciais para cada agrupamento, esta técnica determina, dentre os dados a serem analisados, a que agrupamento cada um vai pertencer [Golub et al. 1999].

- *Redes de Relevância:* visa encontrar e mostrar os pares de genes com forte correlação positiva, ou negativa, e então construir redes com estes pares. Geralmente a força da correlação é proporcional a espessura das linhas usadas para representar as ligações entre os genes, e utilizam-se cores diferentes para representar correlação positiva e negativa [Butte et al. 2000].

- *Visualização:* são análises aplicadas sobre dados muito volumosos ou com característica muito complexa, com a finalidade de conseguir uma melhor compreensão destes dados, identificando para isto alguns padrões que possam estar presentes, ou determinando como estes dados variam. Entre as técnicas de visualização empregadas em Microarranjos estão:

  - *Análise de Componentes Principais (PCA):* técnica analítica muito utilizada como uma técnica de visualização, a qual consegue capturar a variabilidade presente nos dados e representá-la na forma de componentes, sendo que a primeira componente é mais representativa que a segunda, a segunda mais que a terceira, e assim consecutivamente [Hilsenbeck et al. 1999, Raychaudhuri et al. 2000].
  - *Mapas Auto-Organizáveis (SOM):* faz uso de Inteligência Artificial para separar os dados em grupos, sem precisar estipular o número de agrupamentos desejados [Tamayo 1999, Toronen et al. 1999].

- *Classificação dos Dados*: a partir de um conjunto de dados, o objetivo é separar as informações de acordo com classes previamente informadas. O objetivo

principal é, ao surgir um novo dado, conseguir identificar a que classe ele pertence. Nesta tarefa as Máquinas de Suporte Vetorial (SVM) têm sido usadas com sucesso por diversos pesquisadores [Brown et al. 2000][Noble 2004].

Enfim, neste trabalho serão utilizadas diversas técnicas, descritas com mais detalhes no capítulo 3, com o objetivo de determinar, a partir dos dados brutos de Microarranjos, as Redes Regulatórias representativas destes. Na próxima seção serão apresentados alguns conceitos sobre tais redes.

2.3 Redes Regulatórias

A manutenção da vida celular depende de diversas interações que ocorrem em função de fatores intra e extra celulares. Esses fatores abrangem a partir de concentração de certas proteínas até mesmo variações de temperatura e etapas de processos celulares. Estas interações são regidas pelo que se chama de Redes Regulatórias, as quais descrevem a maneira como cada gene atua sobre os produtos de outros genes, ou melhor, o produto resultante da expressão dos genes. Para simplificar, neste trabalho iremos nos referir a esta atuação como sendo a atuação de um gene sobre outro.

Assim, sendo os Microarranjos uma ferramenta para análise da expressão gênica, estes experimentos permitem a observação do estado dos componentes celulares em um dado momento. A partir de experimentos como este, podemos obter diversos tipos de redes de interação, tais como as redes metabólicas, de sinalização, interação proteína-proteína e também as regulatórias. Tais redes são as responsáveis pelo comportamento da célula [Barabási and Oltvai 2004]. No caso das redes regulatórias, elas controlam a operação celular através do nível de expressão gênica, determinando, de maneira dinâmica, quando, e de que forma, cada gene será transcrito em RNA. Cada transcrição leva à síntese de uma determinada proteína através do processo de tradução. Desta forma, pode-se dizer que a Rede Regulatória descreve o funcionamento de um organismo sob uma determinada condição.

Nestas Redes, diversos tipos de interações podem estar presentes. No caso deste trabalho, é estudada a interação entre os genes. Estes tipos de interações, também classificados através de *motivos*, serão apresentados a seguir.

2.3.1 *Motivos* de Interação

O termo *motivo* costuma ser empregado quando se trabalha no nível de análise de seqüências. Em seu trabalho, Shen-Orr *et al* [Shen-Orr et al. 2002] generalizam o termo *motivo* para o nível de redes de interações. Foram chamados de *motivos* os padrões de interconexão presentes em diferentes partes de uma rede, encontrados com uma determinada freqüência.

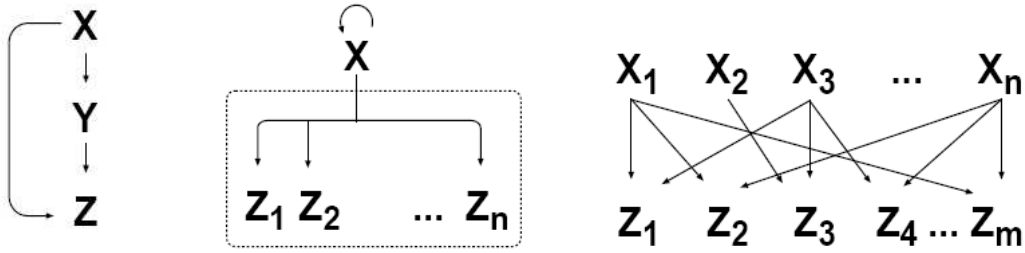
Entre os padrões presentes nas redes, foram encontrados três tipos mais significativos:

- ***Feedforward loop:*** um fator de transcrição X regula um segundo fator de

transcrição Y e os dois em conjunto regulam um ou mais fatores de transcrição Z (Figura 2.4(a)).

- ***Single input module (SIM):*** um único fator de transcrição X regula um conjunto de fatores $Z_1...Z_n$, sempre sob controle do mesmo sinal (todos positivos, ou todos negativos), sendo que X em geral é autoregulador (Figura 2.4(b)).

- ***Dense overlapping regulons (DOR):*** ocorre quando um conjunto de genes $Z_1...Z_m$ são, cada um deles, regulados por uma combinação de alguns fatores de transcrição de um conjunto de fatores $X1...Xn$ (Figura 2.4(c)).



(a) Feedforward loop (b) Single input module (c) Dense overlapping regulons

FIGURA 2.4 – *Alguns motivos de interações encontrados em Redes Regulatórias [Shen-Orr et al. 2002]*

Os motivos apresentados são os mais comuns existentes, mas outros motivos podem estar presentes nos organismos. Determinar os motivos de interação facilita na hora de descrever um circuito responsável por um determinado processo celular, e a análise de tais circuitos é um dos maiores desafios da era pós-genômica [Yeger-Lotem et al. 2004].

A Figura 2.5 mostra a rede de transcrição da *Escherichia coli* (extraída de Shen-Orr *et al* [Shen-Orr et al. 2002]), nesta figura podem ser observadas as diferentes interações existentes entre os genes, bem como seus *motivos*. Fica claro também que diversos genes podem estar regulando um único, de forma positiva, negativa, ou inclusive de ambas as formas, dependendo da situação em que o organismo se encontra. Também vale lembrar que a *E. coli* é um dos organismos mais simples estudado pelo homem, e no entanto, sua rede de transcrição apresenta uma considerada complexidade e é fruto de anos de trabalho e pesquisa.

Assim, se formos considerar a rede de transcrição de outros organismos, a complexidade das relações existentes aumentará consideravelmente. Neste ponto uma metodologia eficiente capaz de levantar indícios de regulação entre os genes se mostra de vital importância frente ao volume de possibilidades de interação gênica existente.

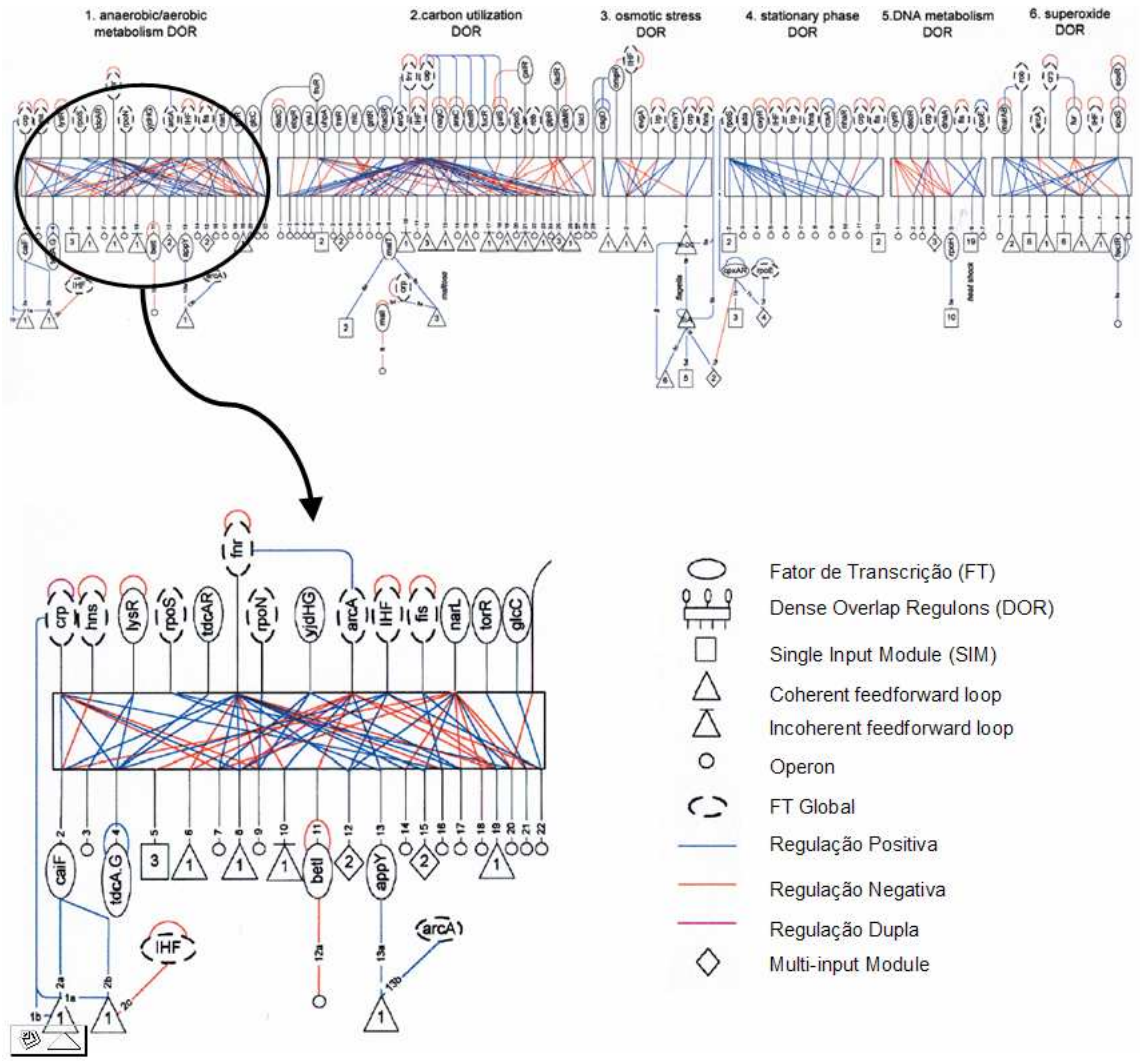


FIGURA 2.5 – *Rede de transcrição da E. coli.*

2.4 Encerramento do Capítulo

Neste capítulo, foram abordados alguns conceitos da Biologia Molecular encontrados na literatura e considerados importantes para a compreensão satisfatória desta dissertação.

No capítulo seguinte, será feita uma exposição dos principais conceitos, pertinentes a este trabalho, sobre Extração de Conhecimento, Mineração de Dados, Redes Neurais e Cadeias de Markov.

Capítulo 3

Conceitos Básicos sobre Técnicas Computacionais

Neste capítulo serão apresentados os principais conceitos sobre as técnicas computacionais utilizadas neste trabalho. Serão apresentados os fundamentos sobre Extração de Conhecimento, enfatizando a etapa de Mineração de Dados, bem como uma introdução às Redes Neurais Artificiais (RNAs), Mapas Auto-Organizáveis (SOM) e Cadeias de Markov.

Primeiramente este capítulo apresentará uma introdução ao processo de Extração de Conhecimento, apresentando suas etapas e comentando brevemente sobre a importância de cada uma delas para o processo como um todo. Esta visão é importante para que o leitor possa se familiarizar com este tipo de metodologia e também o porquê das diversas técnicas aplicadas.

Entre as diversas etapas do processo de Extração de Conhecimento está a de Mineração de Dados. Tal etapa tem grande importância pois é nela em que se usa as diferentes abordagens para atender aos diferentes objetivos da extração de conhecimento, como por exemplo, predição, regressão e classificação. Desta forma foi destinada uma seção deste capítulo para apresentar esta etapa, seus conceitos e metodologias envolvidas.

As etapas de Extração de Conhecimento podem envolver diversas técnicas. No caso deste trabalho há duas em destaque: as Redes Neurais Artificias (RNAs) e os Mapas Auto-Organizáveis (SOMs). As RNAs foram utilizadas como modelo para extrair um modelo para os dados, mais precisamente as Redes Neurais Recorrentes (RNRs). Assim, os fundamentos destas técnicas são apresentadas neste capítulo, bem como dos SOMs, um tipo de RNA de aprendizado não-supervisionado, o qual é aplicado na etapa de pré-processamento.

Por fim, há uma seção dedicada aos conceitos e propriedades das Cadeias de Markov, que é o formalismo adotado para representar o conhecimento extraído, cuja propriedade estocástica intrínseca favorece a adaptação às incertezas presentes no processo.

3.1 Extração de Conhecimento

A cada dia que passa surgem novas técnicas de geração de dados, e como conseqüência disso, as bases de dados têm se tornado muito grandes, inviabilizando a análise manual das informações. Em muitos casos, determinar uma informação pertinente em um banco de dados pode exigir a análise de dezenas de variáveis ao mesmo tempo, sendo assim, existe a necessidade de se automatizar este processo.

Desta forma, chama-se de Extração de Conhecimento o processo de descobrir informações estratégicas ocultas em grandes bases de dados. Neste processo, a partir de dados brutos consegue-se extrair informações de alto-nível[1] com a finalidade principal de auxiliar um especialista a tomar alguma decisão [Fayyad et al. 1996b].

Nesta área, dois termos são muito utilizados: *Extração de Conhecimento* e *Mineração de Dados*. Muitas vezes, estes termos são utilizados inclusive para descrever o mesmo processo. No entando, são coisas diferentes [Fayyad et al. 1996b, Goebel and Gruenwald 1999], considera-se Mineração de Dados como sendo apenas uma etapa do processo de Extração de Conhecimento, onde ocorre a extração de padrões, ou modelos, dos dados observados.

Fayyad [Fayyad et al. 1996a] definiu Extração de Conhecimento como sendo o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis, e compreensíveis, nos dados analisados. Fayyad também definiu cada um dos requisitos:

- **Dados:** conjunto de fatos F que serão analisados.
- **Padrão:** é uma expressão E em uma linguagem L que descreve os fatos em um subconjunto F_E de F.
- **Processo:** quando se fala em processo, está se referindo aos diversos passos necessários à Extração de Conhecimento, os quais envolvem a preparação dos dados, busca por padrões e avaliação dos resultados. Diz-se que estes processos são não triviais por ter um certo grau de autonomia.
- **Validade:** os padrões encontrados precisam ser válidos, e esta validade pode ser obtida comparando os padrões aos dados analisados.
- **Novos:** Dizer que um padrão é novo, é dizer que ele tem informação que ainda não era representada em nenhum outro padrão antes determinado. Padrões redundantes, em geral, são de pouco interesse.
- **Potencialmente úteis:** os padrões determinados devem ter uma aplicação, a fim de serem úteis para alguma inferência desejada sobre os dados.
- **Compreensíveis:** um dos objetivos da Extração de Conhecimento: tornar os padrões compreensíveis para os humanos proporcionando melhor entendimento dos dados em questão.

[1] chama-se de informação de alto-nível aquela informação capaz de ser compreendida pelo homem.

Pela definição, pode ser percebido que o processo de Extração de Conhecimento envolve muitas etapas, que serão abordadas a seguir. Estas etapas estão melhor detalhadas em [Fayyad et al. 1996a, Goebel and Gruenwald 1999, Fayyad et al. 1996b].

3.1.1 Etapas do Processo de Extração de Conhecimento

A Figura 3.1 (adaptada de [Fayyad et al. 1996b]) apresenta as etapas necessárias para o processo de Extração.

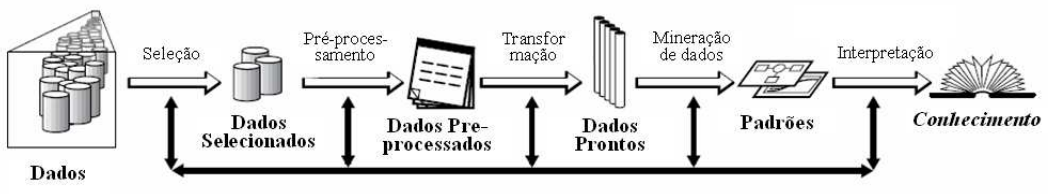


FIGURA 3.1 – *Passos que compõe o processo de Extração de Conhecimento*

Quando tem-se uma base de dados que se deseja analisar, o primeiro passo é conhecer o domínio da aplicação, informando-se sobre os objetivos e, também, sobre que tipo de informação é realmente pertinente ou não. Desta forma pode-se fazer a **seleção** dos dados a serem utilizados com mais segurança. Para o processo nem sempre é necessária a utilização de toda uma base de dados, isto muitas vezes atrapalha a análise, pois agrega uma série de ruídos e informações não significativas. Assim, trabalhar com um subconjunto dos dados, ou diminuir o número de variáveis a se trabalhar, pode ser uma estratégia interessante.

Uma vez que os dados a serem analisados já estão selecionados, a próxima etapa é a **limpeza e pré-processamento**. Nesta etapa são determinadas algumas estratégias para resolver problemas característicos de determinado conjunto de dados. Neste ponto é realizada a remoção de ruídos e *outliers*, bem como aplicam-se estratégias para resolver imperfeições como a falta de alguns dados, ou determinar valores desconhecidos.

O próximo passo é o processo de **transformação**, o qual visa deixar os dados prontos para a etapa de **mineração de dados**. As transformações realizadas sobre os dados visam principalmente reduzir o volume de dados em questão, utilizando somente as informações pertinentes para se determinar os padrões existentes, e também colocar os dados em um formato apropriado para o método de mineração a ser aplicado.

A etapa de mineração é uma das mais importantes de todo o processo. É neste momento que se determina o que será extraído dos dados de acordo com a modelagem aplicada. Esta etapa inclui busca por padrões de interesse, em uma determinada forma. Devido à complexidade desta etapa, ela será melhor explicada na Seção 3.2.

Finalizando, é necessário um processo de **interpretação**, no qual os padrões obtidos são analisados, podendo esta análise proceder tanto matematicamente, como

visualmente. Faz-se a remoção de padrões redundantes e irrelevantes e se traduz estes padrões para informação em uma forma compreensiva pelos usuários.

Enfim, o processo de Extração de Conhecimento tem diversas etapas, e elas foram brevemente mencionadas acima. Cada etapa pode variar de acordo com a realidade do problema que está sendo abordado. Foi considerada como mais importante a etapa de Mineração de Dados, mas nem sempre isto é verdade, isto depende da abordagem e da modelagem feita dos dados. O que pode se afirmar é que a etapa de Mineração poder vir a ser a mais complexa, mas todas as outras etapas são igualmente importantes para o resultado final.

3.2 Mineração de Dados

O ato de minerar os dados pode ser visto como a tarefa de adaptar um modelo, ou determinar padrões, em um conjunto de dados. O modelo em questão terá a função de inferir o conhecimento. Decidir quando o modelo é confiável para se extrair algum conhecimento representativo dos dados em questão, faz parte de todo o processo interativo de Extração de Conhecimento no qual o conhecimento de um especialista é geralmente necessário.

Existe uma grande variedade de algoritmos para mineração de dados oriundos de diversas áreas tais como: estatística, reconhecimento de padrões, aprendizado de máquina e banco de dados. A maioria destes algorítmos pode ser vista como o conjunto de algumas técnicas básicas de cada área, possuindo em geral três componentes:

- **Modelo:** o modelo contém parâmetros que são determinados a partir dos dados. O modelo possui basicamente uma função (por exemplo, classificação ou aproximação) e uma representação (uma aproximação linear ou uma função de probabilidade gaussiana, por exemplo).

- **Critério:** o critério serve para avaliar a pertinência de um modelo, ou conjunto de parâmetros, aos dados analisados. Pode ser algum tipo de função para medir o erro, com algumas adaptações para não especializar o modelo nos dados em questão, e nem dotá-lo de um número de graus de liberdade excessivo.

- **Algoritmo de busca:** um algoritmo que determine a melhor estrutura e parâmetros, uma vez que sejam fornecidos os dados, o modelo e o critério.

Desta forma, pode-se dizer que um algoritmo de mineração de dados consiste na instanciação dos componentes Modelo/Critério/Busca. No entanto os algoritmos possuem grande diferença em cada um dos componentes, mas principalmente nos modelos adotados para realizar a Mineração.

A variação nos componentes utilizados para realizar a mineração, permite que sejam desempenhadas diferentes tarefas dentro do processo de Extração de Conhecimento de acordo com o que se deseja extrair dos dados. A seguir serão abordadas algumas das categorias mais usuais da tarefa de mineração, bem como algumas metodologias utilizadas.

3.2.1 Categorias de Mineração de Dados

Os métodos de mineração possuem diferentes objetivos, sendo comum a aplicação de mais de um método consecutivamente no processo de extração para atingir o objetivo desejado pelo usuário. Os objetivos da mineração podem ser divididos em diversas categorias, e são apresentados a seguir.

- *Processamento dos Dados:* são algoritmos para manipulação dos dados envolvendo técnicas de seleção, filtros, amostragem e transformação dos dados. Visa otimizar o trabalho de análise dos dados.
- *Predição:* realiza a predição do valor de um atributo para um dado fornecido em conjunto com um determinado modelo. Em alguns casos a predição também pode ser utilizada para validar uma nova hipótese.
- *Regressão:* fornecido um conjunto de itens, chama-se de regressão a determinação dos parâmetros para um modelo de aproximação, tornando este capaz de predizer o valor dos atributos para novos itens, ou para itens para os quais não se sabe os valores dos atributos.
- *Classificação:* fornecido um conjunto pré-definido de classes e um conjunto de dados não vistos, determinar a qual classe pertence cada um dos dados.
- *Clusterização:* a partir de um conjunto de itens, particionar este conjunto em diversas classes de tal forma que dados semelhantes estejam agrupados na mesma classe.
- *Associação:* identificar relações existentes entre atributos e itens, tais como a presença de padrões que implicam na presença de um outro padrão. As relações podem ser entre atributos de um mesmo ítem ou de itens diferentes.
- *Visualização:* tem como objetivo fazer com que o conhecimento obtido possa ser facilmente compreendido e interpretado pelos usuários. Uma técnica muito empregada para este fim é a Analise das Componentes Principais (PCA).
- *Análise dos Dados:* visa realizar uma exploração iterativa dos dados sem levar em consideração algum modelo ou conhecimento prévio, a fim de determinar possíveis padrões existentes nos dados.

Assim como existem diversas categorias de Mineração de Dados, também existem diversas metodologias que são amplamente utilizadas. Algumas destas metodologias são abordadas a seguir.

3.2.2 Metodologias de Mineração de Dados

Definir o que é uma metodologia para Mineração de Dados não é uma tarefa simples. Qualquer método capaz de auxiliar na inferência sobre um determinado dado analisado, pode ser considerado como sendo um método de mineração.

Diferentes métodos atendem diferentes propósitos, cada um apresentando vantagens e desvantagens, dependendo de diversos fatores, tais como: a natureza dos dados a serem analisados e o tipo de resposta esperado. No entando, a maioria dos métodos conhecidos, utilizados na mineração, podem ser divididos nos seguintes grupos:

- *Métodos Estatísticos:* métodos focados em verificar hipóteses pré-definidas e adaptar modelos aos dados. Fazem uso de uma abordagem probabilística.

- *Raciocínio Baseado em Casos:* abordagem que tenta resolver um problema baseado em problemas já resolvidos. Analisa-se o problema em questão e verifica-se, dentre os problemas já resolvidos, se existe algum similar este. Existindo, adota-se a mesma solução para o novo problema.

- *Redes Neurais:* modelo que tenta simular o processamento realizado pelo cérebro humano, explicado em maiores detalhes na Seção 3.3. Nesta metodologia, um grande número de neurônios está interconectado através de pesos sinápticos, na tentativa de simular o aprendizado humano. Através de um algoritmo apropriado, a rede "aprende"através de alterações em seus pesos, as quais são realizadas para, na presença de um estímulo (dados), obter a saída desejada.

- *Árvores de Decisão:* são árvores onde cada folha representa um teste ou decisão sobre um ítem dos dados. Dependendo do resultado do teste é escolhido um determinado ramo. Para classificar um dado inicia-se na raíz da árvore, percorrendo seus nodos até atingir uma folha. A folha representa a decisão a ser tomada. As árvores de decisão podem ser consideradas como uma forma especial de um conjunto de regras, caracterizada por sua hierarquia organizacional.

- *Redes Bayeseanas:* são representações gráficas de distribuições de probabilidade baseadas nas ocorrências presentes nos dados. Mais especificamente, são grafos onde os nodos representam atributos e as arestas representam as dependências probabilísticas entre os atributos. Associado a cada nodo, há uma distribuição de probabilidade descrevendo a relação entre o nodo e seus nodos-pais.

- *Algoritmos Genéticos:* pertencentes à classe de algoritmos Evolutivos, são algoritmos de otimização inspirados nos princípios observados na evolução natural. A partir de uma coleção de possíveis soluções que competem umas com as outras, as melhores soluções são selecionadas e combinadas entre si a fim de gerar um novo conjunto de soluções ainda mais eficiente.

- *Conjuntos Difusos:* metodologia própria para trabalhar e processar incerteza. Conjuntos Difusos adaptam-se bem a dados incompletos, com ruídos ou imprecisos, além de apresentar bons resultados quando se deseja uma modelagem inteligente e um controle suave.

Como apresentado, existem diversas metodologias para realização da Mineração de Dados, inclusive diversas delas produzindo um mesmo tipo de resultado. Cabe, a quem for fazer uso destas metodologias, analisar qual se adapta melhor às necessidades.

3.3 Redes Neurais Artificiais - RNAs

Redes Neurais Artificiais (RNAs) são sistemas paralelos distribuídos, compostos por unidades de processamento simples (neurônios) que calculam determinadas funções matemáticas (normalmente não-lineares). O funcionamento destas redes é inspirado em uma estrutura física concebida pela natureza: o cérebro humano. As unidades de processamento são dispostas em uma ou mais camadas, interligadas por um grande número de conexões, que geralmente são unidirecionais. Na maioria dos modelos estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede [Braga et al. 2000].

Entre as características do cérebro humano, que lhe atribuem o comportamento inteligente, as mais importantes a serem simuladas por uma RNA são [Rauber 1998]:

- *Robustez e tolerância a falhas:* a eliminação de alguns neurônios não afeta a funcionalidade global.
- *Capacidade de aprendizagem:* o cérebro é capaz de aprender novas tarefas que nunca foram executadas antes.
- *Processamento de informação incerta:* mesmo que a informação fornecida seja incompleta, afetada por ruído ou parcialmente contraditória, ainda assim um raciocínio correto é possível.
- *Paralelismo:* um imenso número de neurônios está ativo ao mesmo tempo. Não existe restrições que obriguem o processador a executar as instruções seqüencialmente.

3.3.1 Modelo de um neurônio

Um neurônio é uma unidade de processamento de informação, fundamental para a operação de uma rede neural. O modelo neuronal possui três elementos básicos [Haykin 2001]:

- Um conjunto de sinapses ou elos de conexão, cada uma caracterizada por um peso ou força própria. O peso sináptico de um neurônio artificial pode estar em um intervalo que inclui tanto valores negativos quanto valores positivos.
- Um somador para somar os sinais de entrada, ponderados pelas respectivas sinapses do neurônio.
- Uma função de ativação para restringir a amplitude da saída de um neurônio.

O modelo neuronal (Figura 3.2) também inclui um bias. O bias tem o efeito de aumentar ou diminuir a entrada líquida da função de ativação, dependendo se ele é positivo ou negativo, respectivamente.

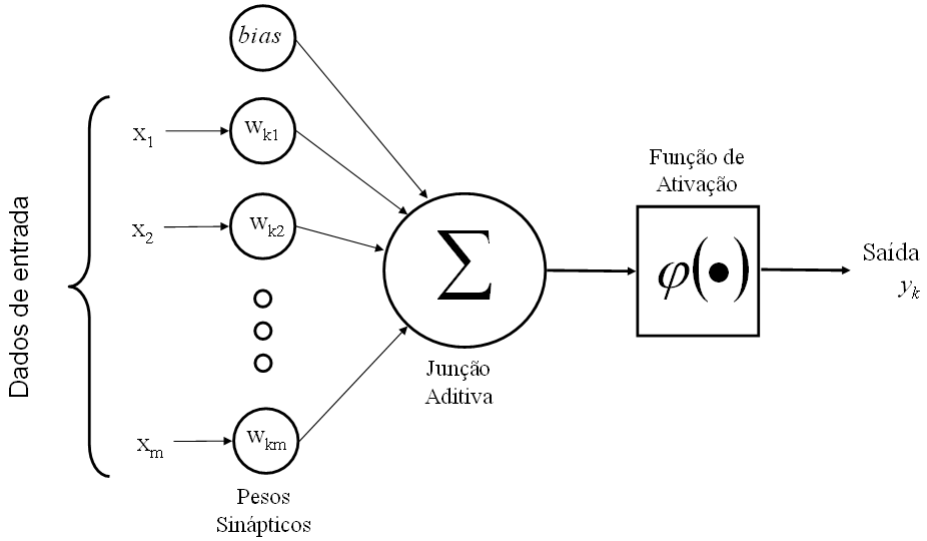


FIGURA 3.2 – *Modelo de um Neurônio*

Em termos matemáticos, podemos descrever um neurônio pela seguinte equação:

$$y_k = \varphi\left(\left(\sum_{j=1}^{m} w_{kj}x_j\right) + b_k\right) \tag{3.1}$$

onde x_1, x_2, ..., x_m são os sinais de entrada; w_{k1}, w_{k2}, ..., w_{km} são os pesos sinápticos do neurônio k; b_k é o *bias*; φ é a função de ativação; e y_k é o sinal de saída do neurônio. Este modelo também é chamado de Modelo McCulloch & Pitts [McCulloch and Pitts 1943] e cabe fazer algumas considerações [Braga et al. 2000]:

1. Redes que seguem este modelo de neurônio, com apenas uma camada, conseguem implementar somente funções linearmente separáveis.

2. Pesos negativos são mais adequados para representar disparos inibidores.

3. O modelo foi proposto com pesos fixos, não-ajustáveis.

3.3.2 Função de Ativação

Considerando $v = \left(\sum_{j=1}^{m} w_{kj}x_j\right) + b_k$, a função de ativação é representada por $\varphi(v)$ e define a saída de um neurônio. Dentre as possíveis funções que poderiam ser utilizadas, quatro se destacam: função linear (Figura 3.3-a), a função rampa (linear por partes) (Figura 3.3-b), a função degrau (step) (Figura 3.3-c) e a função sigmoidal (Figura 3.3-d) [Braga et al. 2000, Haykin 2001, Freeman and Skapura 1992]

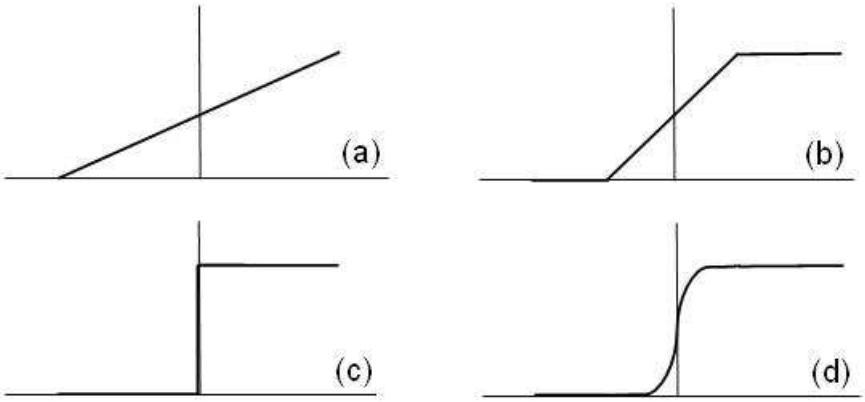


FIGURA 3.3 – *Funções de Ativação*

A função linear é definida por:

$$y = \alpha x + c \tag{3.2}$$

onde α é um número real que define a saída linear para os valores de entrada representados por x a fim de gerar a saída y. Esta função pode sofrer limitações quanto aos seus valores de saída, fazendo com que os valores gerados limitem-se a uma faixa $[\gamma, \lambda]$. Uma função com esta limitação é denominada função rampa.

Na função degrau é definido um valor chamado de limiar, se a entrada for inferior a esse limiar, a saída será 0, se o valor for maior ou igual ao limiar, o valor será 1.

Por último, a função sigmoidal, a forma mais comum de função de ativação utilizada na construção de redes neurais artificiais. Ela é definida como uma função estritamente crescente que exibe um balanceamento adequado entre comportamento linear e não-linear. Um exemplo de função sigmóide é a função logística, definida por:

$$y = \frac{1}{1 + \varepsilon^{\frac{-x}{T}}} \tag{3.3}$$

onde y é o valor de saída, x é a entrada, e T determina a suavidade da curva. Cabe salientar que no limite, quando T tende ao infinito, esta função aproxima-se da função degrau, mas com uma vantagem, a função logística é diferenciável, enquanto a degrau, não. Maiores explicações em Haykin [Haykin 2001].

3.3.3 Arquiteturas de Rede

A definição da arquitetura de uma RNA é um parâmetro importante na sua concepção, uma vez que ela restringe o tipo de problema que pode ser tratado pela

rede. Fazem parte da definição da arquitetura os seguintes parâmetros: número de camadas da rede, número de nodos em cada camada, tipo de conexão entre os nodos e topologia da rede [Braga et al. 2000, Haykin 2001].

- **Redes com camada única.** São ditas redes de camada única por conterem somente uma camada computacional (camada de saída). Possuem uma camada de entrada, mas esta é chamada de camada de entrada de nós fonte, e nela não é realizado nenhum processamento dos dados, logo, não é contada como camada de processamento[Haykin 2001].

- **Redes com camadas múltiplas.** Esta classe de rede diferencia da anterior por possuir camadas ocultas. A função dos neurônios ocultos é intervir entre a entrada externa e a saída da rede a fim de propiciar um melhor aprendizado à rede, adicionando características não-lineares ao modelo. Desta forma, a rede se torna capaz de extrair estatísticas de ordem elevada [Freeman and Skapura 1992].

- **Redes Recorrentes.** Diz-se que uma rede neural é recorrente quando há pelo menos um laço de *realimentação*. A realimentação pode ser de natureza local, quando ocorre no nível de um neurônio, ou de natureza global, se a realimentação engloba uma ou mais camadas completas (Figura 3.4). Os ramos de realimentação envolvem os ditos *operadores de atraso unitário*, os quais propiciam à rede um comportamento dinâmico [Kolen and Kremer 2001].

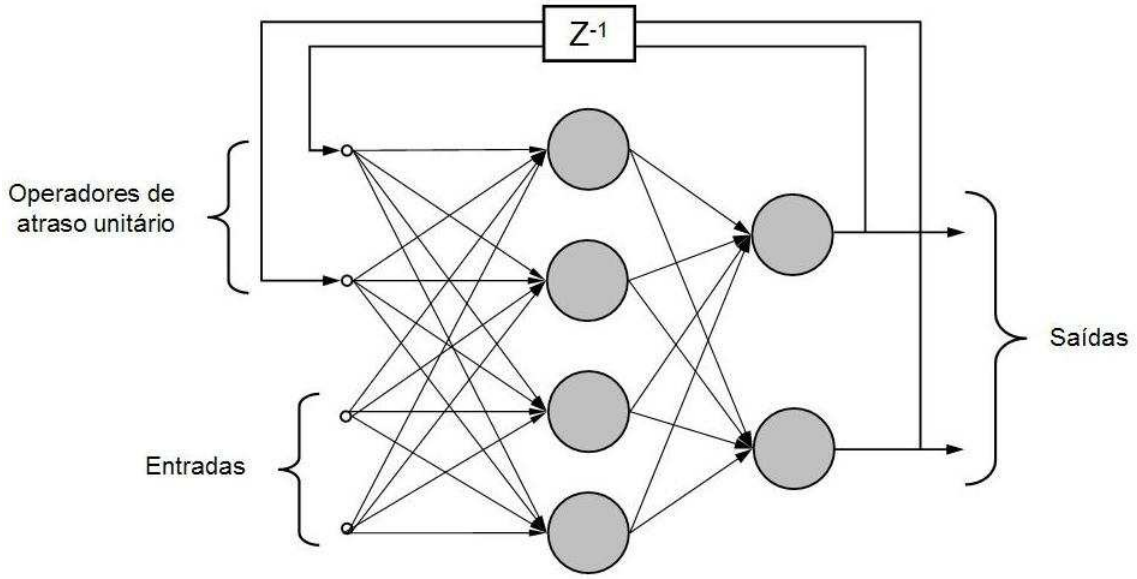


FIGURA 3.4 – *Rede Neural Recorrente*

3.3.4 Aprendizado

Quando se trabalha com redes neurais é necessária uma adaptação dos valores de conexões (pesos sinápticos) para estas redes produzirem os valores esperados, ou seja, para que ela *aprenda* alguma função. Um conjunto de procedimentos bem-definidos para adaptar os parâmetros de uma RNA para que a mesma possa realizar esta adaptação é chamado de *algoritmo de aprendizado*. Mendel e McLaren [Mendel and McLaren 1970] definiram aprendizagem como: *"... o processo pelo qual os parâmetros de uma rede neural são ajustados através de uma forma continuada de estímulo pelo ambiente no qual a rede está operando, sendo o tipo específico de aprendizagem realizada definido pela maneira particular como ocorrem os ajustes realizados nos parâmetros"*.

A forma de aprendizado divide-se em dois tipos principais: aprendizado supervisionado e aprendizado não-supervisionado.

- **Aprendizado Supervisionado**

Esta é a forma mais comum no treinamento das RNAs. O objetivo desta forma de aprendizado é realizar um mapeamento entre os valores de entrada e de saída através da RNA. Nesta metodologia há a presença de um supervisor, o qual indica explicitamente um comportamento para a rede, visando direcionar o processo de treinamento.

O processo consiste em apresentar uma entrada à rede, propagar os sinais e verificar o valor obtido como saída. O valor de saída é comparado com o valor desejado e a diferença entre os dois é considerado como erro para o padrão. Cabe ao supervisor realizar ajustes na rede de forma a minimizar o erro, fazendo com que a rede aprenda o conjunto de dados.

Este tipo de aprendizado tem como vantagem criar diversas funções para um mesmo conjunto de dados, pois, o supervisor consegue direcionar o aprendizado a fim de obter a saída (mapeamento da rede) desejada. Um dos algoritmos mais utilizados para o treinamento supervisionado é o *"error backpropagation"*, descrito por Rumelhart *et al* [Rumelhart et al. 1986].

- **Aprendizado Não-supervisionado**

Como o nome sugere, neste tipo de aprendizado não há a presença de um supervisor para acompanhar o processo de treinamento (Figura 3.5-b). Este aprendizado, assim como o supervisionado, também possui semelhanças com o aprendizado humano, principalmente nos estágios iniciais da vida, na visão e na audição, onde muitos estímulos externos (imagens ou sons) recebem uma resposta própria de cada indivíduo, sem necessariamente o acompanhamento de um supervisor, como por exemplo a reação à dor ou à luz.

Os algoritmos de aprendizado não-supervisionado contam apenas com os dados de entrada da rede. Com a ausência de um supervisor, a rede adapta seus parâmetros de acordo com as estatísticas dos dados de entrada, separando os dados em grupos, da forma mais homogênea possível. Uma das regras mais utilizadas neste

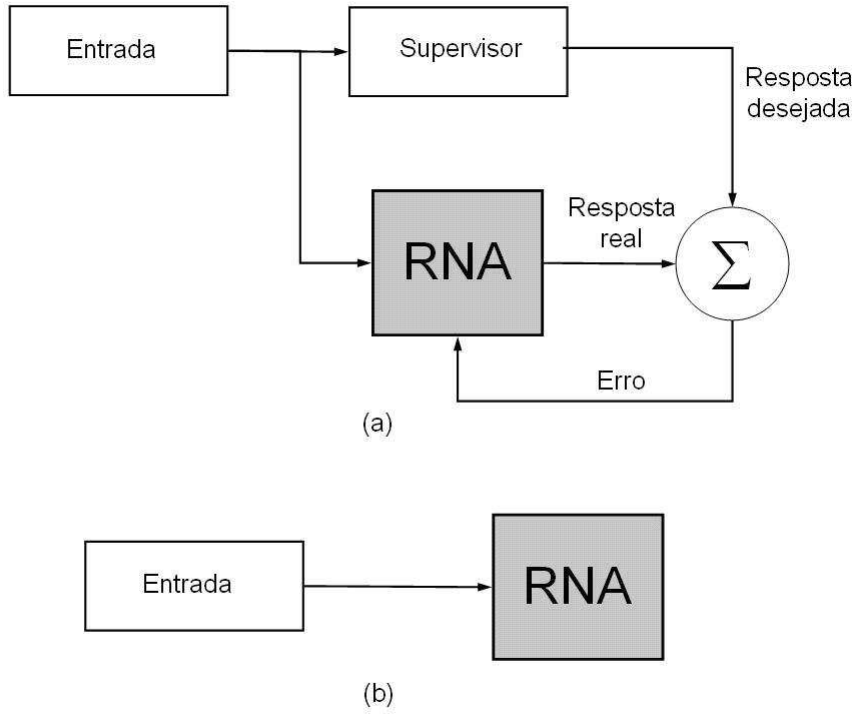


FIGURA 3.5 – *(a) Aprendizado Supervisionado; (b) Aprendizado Não-supervisionado*

aprendizado é o aprendizado competitivo, onde todos os neurônios competem para ver qual o que melhor se aproxima de cada entrada apresentada a rede. O mais próximo é considerado o vencedor, tendo seus valores de pesos atualizados de forma a representar melhor o dado em questão [Kohonen 1997].

Neste trabalho um modelo de rede que usa aprendizado não-supervisionado será utilizado, os Mapas Auto-Organizáveis. Este modelo será brevemente descrito na seção 3.4.

3.3.5 Tipos de problemas solucionados com RNAs

Devido às suas características, as RNAs solucionam diversos tipos de problemas, entre eles destacam-se os de classificação e os de aproximação de função.

Nos problemas de classificação a RNA visa classificar os padrões em diversas classes. Estas classes podem ser conhecidas ou não. Caso sejam conhecidas, o método de aprendizado supervisionado pode ser aplicado. Caso contrário, aplica-se o aprendizado não-supervisionado, que irá encarregar-se de separar os dados visando encontrar as classes e atribuir os padrões das classes existentes.

Quando se aborda o problema de aproximação de função, apresenta-se um valor a rede e faz-se com que ela apresente um resultado o mais próximo possível do desejado. Para isto é necessário que se use uma técnica de aprendizado supervisionado.

3.4 Mapas Auto-Organizáveis

Os Mapas Auto-Organizáveis (SOM) são RNAs que realizam o mapeamento de um espaço de entrada $\mathfrak{R}^n$ para uma matriz bidimensional composta por m neurônios. Cada neurônio i é composto por um vetor de pesos $\vec{m}_i \in \mathfrak{R}^n$. Compara-se um vetor de entrada $\vec{x} \in \mathfrak{R}^n$ com os neurônios, e o melhor neurônio é definido como sendo a "resposta" da rede, sendo a entrada mapeada para este neurônio.

O mapeamento da SOM é considerado por alguns como sendo uma "projeção não-linear" da função densidade de probabilidade de um espaço de mais alta dimensão para um espaço bidimensional [Kohonen et al. 1996].

A determinação do melhor neurônio para uma determinada entrada pode ser realizada utilizando qualquer técnica, mas em geral é considerado o melhor aquele que apresenta a menor distância Euclidiana quando comparado com os valores de entrada. Este neurônio também é chamado, por alguns autores, de neurônio vencedor. Formalmente, a resposta c da rede é determinada da seguinte forma:

$$c = arg\, min_i \left\{ \|\vec{x} - \vec{m}_i\| \right\} \tag{3.4}$$

assim, a entrada $\vec{x}$ é mapeada no neurônio c através do vetor de pesos $\vec{m}_i$.

O processo de aprendizagem do SOM ocorre de acordo com a equação a seguir, onde os pesos iniciais $\vec{m}_i(0)$ podem ser iniciados aleatóriamente.

$$\vec{m}_i\,(t+1) = \vec{m}_i\,(t) + h_{ci}\,(t)\,[\vec{x}\,(t) - \vec{m}_i\,(t)] \tag{3.5}$$

O termo $h_{ci}\,(t)$ apresentado refere-se à função de vizinhança. Durante o aprendizado da SOM, não somente o neurônio considerado como resposta ao dado de entrada sofre ajuste, mas também os neurônios topologicamente próximos a ele. Estes neurônios topologicamente próximos são considerados a vizinhança do neurônio, e o tamanho dessa vizinhança, bem como a forma como ela também será alterada durante o aprendizado, estão definidos nesta função.

A função de vizinhança pode ser determinada de diversas maneiras, a mais simples é determinar um perímetro ao redor da resposta, e atualizar os pesos de todos os neurônios dentro desse perímetro. Kernels deste tipo são chamados de bolha. No entanto, kernels mais complexos são amplamente empregados, como o baseado na função Gaussiana:

$$h_{ci} = \alpha\,(t)\,.exp\left(-\frac{\|r_c - r_i\|^2}{2\varphi^2\,(t)}\right), \tag{3.6}$$

onde $\alpha(t)$ é uma função para a taxa de aprendizagem, $\|r_c - r_i\|$ é a distância entre a resposta c e o neurônio i, $\varphi(t)$ determina a amplitude da vizinhança. Tanto $\alpha(t)$ como $\varphi(t)$ são funções decrescentes no tempo, e sua forma exata não é determinante para o sistema. É aconselhável a utilização destas funções decrescentes, e um treinamento exaustivo para atingir um bom balanceamento nos resultados, entretanto, se torna difícil determinar *a priori* uma regra para estipular a configuração

ideal para cada rede, isto deve ser feito experimentalmente, analisando diferentes configurações.

3.5 Cadeias de Markov

Cadeias de Markov são utilizadas em problemas de tomada de decisão nos quais existam alguma incerteza associada. Esta incerteza pode ser atribuída a variações presentes nos dados, as quais são comuns quando se trabalha com dados oriundos de fenômenos naturais, como os níveis de expressão gênica analisados neste trabalho. Esta variação necessita ser quantitativamente tratada, sendo uma solução incorporá-la em um modelo matemático. No entanto, fenômenos naturais costumam apresentar certo grau de regularidade, desta forma, sua variação pode ser descrita por um modelo probabilístico [Hillier and Lieberman 1995].

Processos envolvendo modelos probabilísticos são chamados de **processos estocásticos**, e entre estes processos, encontram-se as Cadeias de Markov, que apresentam uma propriedade bem característica: as probabilidades de como o processo irá evoluir dependem somente do estado presente, e não dos estados anteriores. A seguir serão descritos em maiores detalhes os processos estocásticos e as Cadeias de Markov.

3.5.1 Processos Estocásticos

Um processo estocástico é definido como sendo uma coleção de variáveis aleatórias indexadas, X_t, onde o índice t é definido em um conjunto T. Freqüentemente T é tido como o conjunto dos números inteiros não-negativos, e X_t representa uma característica de interesse, mensurável, no tempo t.

Quando considera-se o comportamente de um sistema operando durante um período de tempo, considera-se t como sendo pontos específicos no tempo, chamados de 0, 1, ... Em cada ponto no tempo o sistema encontra-se em uma categoria ou *estado* de um número finito de categorias ou estados mutuamente exclusivos e exaustivos, chamados de 0, 1, ..., M. Os pontos no tempo podem ser igualmente espaçados, ou o espaço entre eles pode depender do comportamento geral do sistema físico no qual o processo estocástico está *inserido*. Embora os estados possam constituir uma caracterização tanto qualitativa quanto quantitativa do sistema, não existe nenhuma perda de generalidade pela denominação numérica 0, 1, ..., M, a qual será usada para denotar os possíveis estados do sistema. Assim, a representação matemática do estado do sistema é a de um processo estocástico $\{X_t\}$, onde as variáveis aleatórias são observadas em $t = 0, 1, 2, ...$, e onde cada variável aleatória pode assumir qualquer um dos $(M + 1)$ números inteiros 0, 1, 2, ..., M, que caracterizam os estados do processo [Karlin and Taylor 1975].

3.5.2 Cadeias de Markov: Definição

Para um processo estocástico ser considerado uma Cadeia de Markov a probabilidade condidional de qualquer "evento" futuro, dado qualquer "evento" passado e

o estado presente $X_t = i$, é independente do evento passado e depende somente do estado presente do processo. Esta característica é chamada de *propriedade markoviana* [Isaacson and Madsen 1976].

As probabilidades condicionais $P\{X_{t+1} = j | X_t = i\}$ são chamadas de probabilidade de transição. Se, para cada i e j,

$$P\{X_{t+1} = j | X_t = i\} = P\{X_1 = j | X_0 = i\}, \forall\, t = 0,\ 1,\ ..., \tag{3.7}$$

então, as probabilidades de transição são ditas *estacionárias* e normalmente denominadas por P_{ij}. Ter probabilidades de transição estacionária implica em elas não mudarem no tempo. A existência de tais probabilidades de transição também implica em, para cada i, j e n ($n = 0,\ 1,\ 2,\ ...$),

$$P\{X_{t+n} = j | X_t = i\} = P\{X_n = j | X_0 = i\} \tag{3.8}$$

para todo $t = 0,\ 1,\ ...$ Estas probabilidades são também denotadas por $p_{ij}^{(n)}$ e, é exatamente a probabilidade condicional de que a variável aleatória X, começando no estado i, esteja no estado j depois de exatamente n etapas (unidades de tempo).

Todas as probabilidades condicionais devem satisfazer as seguintes propriedades [Imperatore and Bentjerodt 1970]:

$$p_{ij}^{(n)} \geq 0 \text{ para todo } i \text{ e } j, \text{ e } n = 0, 1, 2, ...$$

$$\sum_{j=0}^{M} p_{ij}^{(n)} = 1, \text{ para todo } i \text{ e } n = 0, 1, 2, ...$$

As probabilidades condicionais também podem ser representadas convenientemente na forma de matriz

$$P^{(n)} = \begin{array}{c|ccc} \text{Estado} & 0 & 1 & M \\ \hline 0 & p_{00}^{(n)} & \cdots & p_{0M}^{(n)} \\ 1 & & & \\ 2 & \vdots & & \vdots \\ \vdots & & & \\ M & p_{M0}^{(n)} & \cdots & p_{MM}^{(n)} \end{array}, \text{ para } n = 0, 1, 2, ...$$

Assim, pode-se definir uma *cadeia de Markov de estado finito* como sendo um processo estocástico $\{X_t\}(t = 0, 1, ...)$ com as seguintes características:

1. Um número finito de estados,
2. A propriedade markoviana,
3. Probabilidades de transição estacionárias,
4. Um conjunto de probabilidades iniciais $P\{X_0 = i\}$ para todo i.

3.5.3 Equações de Chapman-Kolmogorov

As probabilidades de transição $\left(p_{ij}^{(n)}\right)$ nos mostram, depois de n períodos, qual a probabilidade do processo, estando no estado i, transicionar para o estado j. No entanto, não foi apresentado um método para determinação destas probabilidades. Esta pode ser feita através das *equações de Chapman-Kolmogorov* [Hillier and Lieberman 1995]:

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(v)} p_{kj}^{(n-v)} \tag{3.9}$$

Segundo a Equação 3.9, para ir de um estado i para o estado j em n etapas, o processo estará em algum estado k depois de exatamente v etapas. Desta forma, $p_{ik}^{(v)} p_{kj}^{(n-v)}$ é a probabilidade condicional do processo, iniciando no estado i, ir para o estado k, após v etapas e, então, para o estado j em $n - v$ etapas. Assim, as probabilidades de transição de n etapas podem ser obtidas recursivamente a partir das probabilidades de transição de uma etapa.

Considerando a notação matricial, pode-se dizer que:

$$P^{(n)} = P^n = P^v \cdot P^{n-v} = P \cdot P^{n-1} \tag{3.10}$$

ou seja, a matriz de probabilidades de transição de n estados pode ser obtida calculando-se a n-ésima potência da matriz de transição de uma etapa.

3.5.4 Tempo da Primeira Passagem e de Recorrência

Além de definir as probabilidades de transições entre os estados de uma cadeia, algumas vezes também se deseja saber o número de transições feitas por um processo para ir de um estado i para o j. Quando trata-se do tempo em que estas transições ocorreram pela primeira vez, este tempo é chamado de *tempo da primeira passagem* indo do estado i para o j. Entretanto há um caso especial, quando $j = i$, neste caso o tempo da primeira passagem é o número de transições para o processo retornar ao estado inicial i. Este tempo é então chamado de *tempo de recorrência* para o estado i [Hillier and Lieberman 1995].

Os tempos das primeiras passagens são variáveis aleatórias e têm distribuições de probabilidade associadas a elas. Tais distribuições dependem das probabilidades de transição do processo. Denotando a probabilidade de que o tempo da primeira passagem do estado i para o j seja igual a n por $f_{ij}^{(n)}$, as seguintes relações mostram como as probabilidades de primeira passagem de i para j em n etapas podem ser calculadas recursivamente a partir das probabilidades de transição de uma etapa.

$$\begin{aligned}
f_{ij}^{(1)} &= p_{ij}^{(1)} = p_{ij}, \\
f_{ij}^{(2)} &= p_{ij}^{(2)} - f_{ij}^{(1)} p_{jj}, \\
&\vdots \\
f_{ij}^{(n)} &= p_{ij}^{(n)} - f_{ij}^{(1)} p_{jj}^{(n-1)} - f_{ij}^{(2)} p_{jj}^{(n-2)} \ldots - f_{ij}^{(n-1)} p_{jj}.
\end{aligned} \tag{3.11}$$

Para i e j fixos, os $f_{ij}^{(n)}$ são números não-negativos tais que:

$$\sum_{n=1}^{\infty} f_{ij}^{(n)} \leq 1. \tag{3.12}$$

Há casos em que esta soma pode ser estritamente menor que 1, indicando que um processo inicialmente no estado i pode nunca chegar ao estado j. Considerando $i = j$, pode-se dizer que se o somatório for menor que 1, há uma probabilidade positiva de o processo, uma vez tendo passado pelo estado i, não retornar mais a este estado, sendo o estado i chamado de *estado transiente*.

Considerando ainda $i = j$ diz-se que quando a soma da Equação 3.12 for igual a 1, o estado i é um *estado recorrente*, pois há uma probabilidade 1 de o processo partindo do estado i retornar a ele mesmo. No entanto, se a condição de recorrência for satisfeita para $n = 1$, então este estado é dito ser um estado de *absorção*.

Quanto existe um estado de absorção k, uma vez que este estado seja visitado o processo não mais sairá deste. Assim, a probabilidade da primeira passagem de i para k é chamada de probabilidade de absorção em k, partindo de i [Isaacson and Madsen 1976].

3.5.5 Propriedades dos Estados de uma Cadeia de Markov

A Seção anterior definiu dois tipos de estados de uma Cadeia de Markov, os recorrentes e os transientes. No entanto, há outras propriedades e conceitos importantes a serem definidos [Hillier and Lieberman 1995]:

- **Acessibilidade:**

Sendo $p_{ij}^{(n)}$ a probabilidade condicional do processo, iniciando no estado i, estar no estado j depois de n etapas, diz-se então que o estado j é *acessível* a partir do estado i se $p_{ij}^{(n)} > 0$ para algum $n \geq 0$.

- **Comunicação:**

Quando um estado j for acessível a partir de um estado i e vice-versa, eles são ditos *comunicados*. Uma Cadeia de Markov na qual todos os estados se comunicam é dita ser uma Cadeia *irredutível*.

- **Periodicidade:**

Em uma Cadeia de Markov, diz-se que um estado i tem um período t ($t > 1$), se $p_{ij}^{(n)} = 0$, sempre que n não for divisível por t, e t for o menor número inteiro com esta propriedade. Um exemplo disto é um processo que somente pode entrar no estado i nos tempos 0, 3, 6, ..., neste caso este estado tem período 3. Se o processo puder estar em um estado em tempos consecutivos, s e $s + 1$, o estado é dito ter período 1 e é chamado de *aperiódico*.

Há ainda um caso especial entre os estados aperiódicos, quando além de aperiódico ele também for recorrente, então passa a se chamar de estado *ergódico* [Isaacson and Madsen 1976].

3.6 Encerramento do Capítulo

Este capítulo apresentou alguns dos principais conceitos encontrados na literatura sobre a metodologia de Extração de Conhecimento, Redes Neurais Artificiais e Cadeias de Markov, cuja exposição é necessária para uma leitura satisfatória deste trabalho.

Foram apresentados alguns conceitos sobre Extração de Conhecimento, apresentando as principais etapas do processo. Foi dedicado uma Seção para tratar, em maiores detalhes, a etapa de Mineração, uma vez que esta é considerada a etapa mais complexa de todo o processo.

Sobre RNAs foram vistos seus conceitos básicos, uma noção sobre arquiteturas de redes, técnicas aplicadas no seu aprendizado, bem como os tipos de problemas a que se aplicam. Foi abordado em maiores detalhes os SOMs, importantes para a aplicação nesta metodologia, na fase de pré-processamento dos dados.

Finalmente foram apresentados alguns conceitos sobre Cadeias de Markov e Processos Estocásticos, sendo que estes farão parte da metodologia e da representação do conhecimento utilizado neste trabalho.

No próximo Capítulo será feita uma análise sobre alguns dos principais trabalhos encontrados no meio científico, referentes, de forma direta, ou indireta, aos temas abordados nesta dissertação.

Capítulo 4

Revisão Bibliográfica

Neste capítulo serão apresentados alguns trabalhos vinculados à manipulação de dados provenientes de Microarranjos e, também, à Extração de Conhecimento.

Primeiramente abordam-se trabalhos sobre processamento de matrizes de expressão gênica para determinar informações relevantes nos dados, bem como determinar os clusters presentes. Trabalhos de classificação de dados de Microarranjos também são abordados.

Em seguida são apresentados alguns trabalhos sobre Extração de Conhecimento a partir de dados de Microarranjos, citando os modelos mais utilizados para esta tarefa. No entanto, não foram encontrados trabalhos que utilizem RNRs para Extração de Conhecimento de Microarranjos, assim, trabalhos utilizando RNRs são apresentados em uma seção a parte.

4.1 Microarranjos

A técnica de Microarranjos é considerada uma técnica nova. Vários trabalhos têm sido desenvolvidos ainda procurando como melhor explicar esta técnica, abordando detalhes de seu funcionamento. Duggan *et al* [Duggan et al. 1999] apresenta, em detalhes, como utilizar esta metodologia na análise da expressão gênica, apresentando os fundamentos da técnica e suas etapas. Butte [Butte 2002] em um trabalho semelhante, relaciona também as bases de dados disponíveis e suas características, tais como os organismos analisados, quantidade de genes envolvidos e o número de amostras no tempo. Por fim, entre os trabalhos introdutórios, Kuo *et al* [Kuo et al. 2004] faz uma abordagem voltada à aplicação de técnicas de Aprendizado de Máquina para extrair informações pertinentes dos dados, citando os modelos aplicados à análise de Microarranjos.

4.2 Seleção de Características

Um dos desafios encontrados ao trabalhar com Microarranjos é o grande número de genes em relação ao número de amostras da maioria dos experimentos. Em um conjunto de dados obtidos em um experimento, apenas parte deles

são pertinentes para análise, o resto pode ser considerado como ruído proveniente dos próprios Microarranjos. Tendo em mente a eliminação destes ruídos, diversos autores propuseram técnicas novas de pré-processamento.

Chuang *et al* [Chuang et al. 2003, Chuang et al. 2004], para solucionar o problema de seleção de genes, propõem classificar os genes dos experimentos. Os dados são estatisticamente analisados através de uma abordagem denominada RAC (Rank and Combination Analysis), a qual consiste de duas etapas. A primeira ordena os genes de acordo com a importância de sua informação. Para medir esta importância, várias medidas podem ser aplicadas, a abordagem RAC realiza uma ordenação para cada método escolhido (entre as medidas utilizadas estão: variância, desvio padrão, correlação e probabilidades). A segunda, consiste em combinar os resultados já obtidos, a fim de determinar uma ordenação final para o conjunto de genes. Os autores construiram um *framework* para este tipo de análise e observaram melhores resultados quando combinados diversos métodos.

Uma abordagem diferente, mas com a mesma intenção, foi feita por Holter *et al* [Holter et al. 2000], onde eles discutem a utilização de SVD (Singular Value Decomposition) [Bjornsson and Venegas 1997, Watkins 1991] para representação da matriz de expressão gênica de um experimento. Segundo eles, um conjunto complexo de expressões gênicas pode ser representado por um pequeno número de modos, obtidos pela SVD, os quais capturam os padrões temporais de alterações na expressão gênica. Foram analisados Microarranjos do ciclo celular CDC-15, esporulação, e fibroblastos. Nos 3 tipos de dados observou-se que apenas as duas primeiras características determinadas pela SVD já mostravam uma boa representação dos dados.

Hitoshi Iba, em seus trabalhos [Ando and Iba 2004, Paul and Iba 2005], vem utilizando para seleção de características uma metodologia baseada em Computação Evolutiva, aplicando algoritmos genéticos nesta tarefa. Seu trabalho diferencia-se por não utilizar operadores de mutação e recombinação e, sim, gerar as novas soluções baseado na distribuição de probabilidades das soluções selecionadas da geração anterior (PMBGA - Probabilistic Model Building Genetic Algorithm).

Entretanto, muitos autores optam por realizar a seleção de características através da determinação dos clusters presentes nos dados. Vários são os métodos aplicados, sendo difícil determinar, *a priori*, qual produz os melhores resultados. Sendo assim, a próxima seção é dedicada a alguns trabalhos sobre o assunto.

4.3 Clusterização

O objetivo da clusterização é subdividir o conjunto de dados (no caso, séries temporais representando o nível de expressão de um gene) de forma a agrupar expressões semelhantes em um mesmo grupo.

Um trabalho considerado clássico na área foi o de Eisen *et al* [Eisen et al. 1998]. Eisen clusterizou dados de Microarranjos de *Saccharomyces cerevisiae* utilizando clusterização hierárquica. Como resultado da técnica, é gerado um dendograma de expressões gênicas. Entretanto, determinar o número de clusters

presentes não é uma tarefa simples, pois depende de uma valor empírico de semelhança que deve ser arbitrado. Mas Eisen pode fazer importantes observações sobre a análise, entre elas a eliminação de dados redundantes propiciada e, também, a tendência de genes de função semelhantes estarem presentes em um mesmo grupo.

Outro método utilizado para clusterização é o *K-means*. Este método utiliza k centróides para particionar os dados. Golub *et al* [Golub et al. 1999] fizeram uso desta técnica em seu trabalho, no entanto, ela apresenta um problema da própria metodologia, necessitar do número de centróides e a posição dos mesmos. Isto torna a técnica sensível à inicialização dos seus parâmetros e, em geral, determinar o número de centróides não é uma tarefa simples, exigindo repetidos experimentos até identificar a configuração adequada para os dados. Um método bastante semelhante ao *k-means*, mas menos exigente quanto aos parâmetros são os Mapas Auto-organizáveis (SOMs).

Segundo Slonim [Slonim 2002], uma das vantagens da utilização de SOMs, em relação ao *k-means* é poder identificar as relações entre os clusters. Desta forma, SOMs têm sido utilizados por diversos autores [Tamayo 1999, Toronen et al. 1999, Oja et al. 2002, Nikkila et al. 2002] para a determinação de clusters em dados de Microarranjos. Foi evidenciado que através da análise do SOM, já treinado, pode-se chegar a muitas conclusões pertinentes quanto aos clusteres presentes. Törönen *et al* [Toronen et al. 1999] em seu trabalho, evidenciou a praticidade de utilização de SOMs na análise de dados de Microarranjos, os quais permitem facilmente a visualização da quantidade de clusters presentes nos dados.

Em um trabalho posterior, Nikkilä *et al* [Nikkila et al. 2002] aprimorou a utilização de SOMs para não apenas visualizar os clusters presentes, mas, também, determiná-los. A determinação dos clusters foi realizada com uso da *U-matrix*, a qual, de modo gráfico, mostra as distâncias entre os neurônios da rede, sendo que os locais onde estas distâncias são menores indica a presença de clusters (maiores detalhes sobre o método ver Ultsh [Ultsch 1993]). Neste trabalho não foi possível determinar com exatidão os clusters, mas quanto à visualização dos dados a SOM foi mais eficiente que a Clusterização Hierárquica.

Oja *et al* [Oja et al. 2002] também domonstram a utilização de SOMs para a exploração de clusters, abordando como interpretar os resultados da rede e identificar os clusters, no entando, uma de suas mais importantes contribuições foi demonstrar que as mesmas inferências sobre os dados realizadas utilizando Clusterização Hierárquica, podem também ser feitas utilizando SOMs.

Segundo D'haeseleer [D'haeseleer 2005], novos métodos de clusterização são facilmente criados. Entre os novos métodos já desenvolvidos, Ben-Dor & Yakhini [Ben-Dor et al. 1999] modificam o tradicional método *k-means* a fim de otimizá-lo para a análise de Microarranjos com dados de diversas condições. Definem um modelo estocástico para a entrada, também utilizado como medida de similaridade. Yeung & Ruzzo [Yeung and Ruzzo 2001] propõem clusterizar os dados utilizando, para isto, as Componentes Principais extraídas por PCA. Além destas, existem outras propostas de metodologias, como as encontradas em [Lukashin and Fuchs 2001, Bar-Joseph et al. 2002, Zhao and Zaki 2005].

A próxima seção apresenta a compilação de alguns trabalhos realizados para

classificar expressões gênicas de Microarranjos.

4.4 Classificação

Algoritmos de aprendizado de máquina, em geral, trabalham com um grande número de amostras e um número limitado de variáveis, no entanto ocorre o inverso quando classifica-se expressão gênica. Microarranjos apresentam um grande número de variáveis com reduzido número de amostras, tornando-se desta forma um desafio para as técnicas tradicionais de classificação [Slonim 2002].

Considerando as técnicas de aprendizado de máquina, os métodos de classificação estão entre os supervisionados, onde existe uma fase de treinamento com um conjunto de dados conhecidos, e uma fase de teste, onde se classifica um dado ainda não visto em alguma das classes do treinamento. Um dos métodos mais simples de classificação, mas nem por isso menos poderoso, é o de vizinhos mais próximos (*Nearest Neighbors*), onde classifica-se a amostra x como pertencente àquela classe que for considerada mais próxima entre as k classes de treinamento, de acordo com a medida de similaridade adotada. Apesar de simples, vários são os autores que obtiveram sucesso com a utilização desta técnica [Golub et al. 1999, Ramaswamy et al. 2001, Pomeroy et al. 2002].

Liang & Kelemen [Liang and Kelemen 2002] utilizaram Redes Neurais Recorrentes para classificar expressão gênica. Em seu experimento eles enfatizam a necessidade do pré-processamento antes do treinamento da rede. Segundo resultados apresentados, a RNR mostrou-se superior a outras técnicas, entre elas o *K-means*, o SOM e até mesmo a SVM. Mas Liang & Kelemen admitem que estes resultados podem ser por causa dos dados escolhidos, além da técnica tornar-se inapropriada para um grande número de classes, principalmente se as diferenças entre as classess forem sutís.

Algoritmos evolutivos também são utilizados para classificação de expressão gênica. Em geral utiliza-se técnicas baseadas em Algoritmos Genéticos, mediante algumas modificações, com o intuito de selecionar um número reduzido de genes, e então aplicar uma outra técnica em conjunto para classificar o gene [Wahde and Szallasi 2006]. Em um trabalho recente, Ando & Iba [Ando and Iba 2004] analisaram o uso de Algoritmos Genéticos em conjunto com técnicas de aprendizado de máquina (*K-means*, SVM e Naive Bayes) para classificar expressões gênicas. Segundo seus resultados os classificadores apresentaram melhores resultados quando aplicados em conjunto com Algoritmos Genéticos.

Não há um consenso sobre qual técnica melhor se adapta a análise de dados de Microarranjos, no entanto, quanto à classificação de expressão gênica, muitos autores concordam que para tal tafera as SVMs são uma escolha segura [Butte 2002, Brown et al. 2000, Slonim 2002]. A diferença das SVMs em relação a outras técnicas é a facilidade que ela apresenta para lidar com classificação de dados quando, no conjunto de treinamento há um grande número de exemplos negativos para um pequeno número de positivos. Entretanto, SVMs apresentam como desvantagem a difícil interpretação dos resultados.

4.5 Extração de Conhecimento de Microarranjos

O grande volume de dados de um Microarranjo, nem sempre significa um grande volume de informações. Além disto, seu formato dificulta a interpretação das informações. Desta forma, um dos desafios na análise de Microarranjos é conseguir representar a informação presente na forma de Redes Biológicas. Segundo Styczynski & Stephanopoulos [Styczynski and Stephanopoulos 2005] várias técnicas são aplicadas para "decifrar", tanto quantitativamente quanto qualitativamente, a arquitetura das redes.

Para este fim, vários modelos têm sido utilizados. Os mais populares serão apresentados nas seções seguintes.

4.5.1 Modelos Contínuos

Alguns autores abordam o problema utilizando modelos contínuos. Chen *et al* [Chen et al. 1999] foram um dos primeiros a realizar este tipo de modelagem com dados de Microarranjos. Eles propuseram uma modelagem utilizando equações diferenciais. De acordo com a complexidade dada ao problema na hora da modelagem, pode-se ter equações lineares ou não lineares [Styczynski and Stephanopoulos 2005].

Apesar de estes métodos conseguirem uma boa descrição dos dados, determinar os parâmetros do modelo se torna um problema devido à complexidade matemática que o problema adquire. Desta forma, alguns trabalhos voltam seus esforços para determinar uma forma de encontrar os parâmetros do modelo, e não determinar o modelo em si.

Um modelo bastante conhecido para descrever redes bioquímicas é o *S-system* que descreve o sistema por um conjunto de equações diferenciais não-lineares. Este modelo foi utilizado por Sakamoto & Iba [Sakamoto and Iba 2001] e Noman & Iba [Noman and Iba 2005] para determinar redes de regulação gênica usando técnicas de computação evolutiva na estimação dos parâmetros do sistema.

4.5.2 Análise de Correlação

Butte *et al* [Butte and Kohane 2000, Butte et al. 2000] utilizam Redes de Relevância para determinar a interação entre os genes. Esta metodologia analisa a correlação entre os dados, determinando a relação entre eles. Além de mostrar, na forma de uma rede, a relação entre os dados, este método também pode ser utilizado para clusterização. No caso de aplicar esta metodologia para clusterização, a vantagem, em relação às demais técnicas, é o fato de agrupar as expressões gênicas de acordo com sua correlação, independente de serem negativas ou positivas, formando, desta forma, clusters biologicamente representativos.

4.5.3 Modelos Estocásticos

Modelos probabilísticos são utilizados por diversos autores na modelagem de redes regulatórias. Segundo Styczynski & Stephanopoulos [Styczynski and Stephanopoulos 2005] esta modelagem possui vantagens significa-

tivas, como a fácil representação da incerteza e o tratamento de dados faltantes. Como desvantagem, apresenta a grande exigência computacional de seu aprendizado, permitindo apenas a modelagem de pequenas redes, bem inferiores às redes reais.

Murphy & Mian [Murphy and Mian 1999] propuseram o modelo de Redes Bayesianas Dinâmicas (DBN) para modelar dados de Microarranjos. Tal modelo reune características de Redes Bayesianas e Cadeias de Markov. Uma das vantagens desta técnica é a possibilidade de inserção de conhecimento prévio na modelagem. Esta técnica foi utilizada por Ong *et al* [Ong et al. 2002] para analisar séries temporais de Microarranjos, identificando nos dados as relações entre genes. Kim *et al* [Kim et al. 2003] e Perrin *et al* [Perrin et al. 2003] também utilizaram DBNs na inferência de redes genéticas.

Apesar de identificar as relações entre os genes, as DBNs apresentam baixa precisão e exigem muito tempo de processamento. Para minimizar estes problemas, Zou & Conzen [Zou and Conzen 2005] propuseram a limitação dos possíveis genes reguladores, minimizando o tempo de processamento e aumentando a precisão do método. Zhou *et al* [Zhou et al. 2004] também utilizaram uma metodologia semelhante utilizando, em conjunto, Redes Bayeseanas e MCMC (Markov Chain Monte Carlo) para extrair as relações entre os genes.

Entretanto, não existem trabalhos que utilizem RNAs ou RNRs para Extração de Conhecimento de Microarranjos. Na próxima seção é apresentada uma compilação dos trabalhos mais significativos nesta área.

4.6 Extração de Conhecimento de RNRs

Extração de Conhecimento de RNRs baseia-se na extração de regras das mesmas. Estas regras são utilizadas para construir um modelo formal que represente a computação realizada pelas redes [Jacobsson 2005]. Entre os formalismos utilizados estão as máquinas de estados. Elman [Elman 1990], em seu artigo, apresentou o seu modelo recorrente, comparando as ativações internas aos estados de Máquinas de Estados.

Em 1995, Siegelmann & Sontag [Siegelmann and Sontag 1992] mostraram que RNRs são equivalentes a Máquinas de Turing, logo, possuem a capacidade de computar qualquer função computável. Com isto, tem-se uma poderosa ferramenta computacional e, frente a possibilidade de interpretar o seu conhecimento representando sua computação através de Máquinas de Estados, um bom modelo para Extração de Conhecimento.

Diversos trabalhos têm apresentado formas de extração de regras a partir de RNRs [Cechin 1997, Andrews and Geva 2002, Zhou 2004, Vahed and Omlin 2004, Liu et al. 2004, Jacobsson 2005, Pechmann and Cechin 2005]. Resumidamente, os passos para esta extração são:

- discretização do espaço de estados contínuo das RNRs;
- geração dos estados e saídas através da alimentação da RNR com os dados de entrada;

- construção da base de regras com as transições de estados observadas;
- e, por fim, minimização da base de regras.

A discretização do espaço de estados consiste em mapear o espaço de estados contínuo da RNR (microestados), para os estados correspondentes à máquina resultante (macroestados). Este mapeamento é considerado crítico para os algoritmos de extração de regras de RNRs. Tal processo é denominado por "quantização" do espaço de estados.

Várias técnicas para quantização podem ser utilizadas. Servan-Schreiber *et al* [Servan-Schreiber et al. 1989, Servan-Schreiber et al. 1991] utilizaram Clusterização Hierárquica para determinar os estados de uma RNR treinada com strings geradas por uma pequena máquina de estados finita. Nestes experimentos os clusters identificados puderam ser relacionados aos estados da máquina que o gerou, o que representou a conexão entre as RNRs e as máquinas de estados.

Entretanto, nem sempre há uma relação de "um para um" entre os clusters encontrados nas redes e os estados das máquinas. Servan-Schreiber [Servan-Schreiber et al. 1991] observaram esta característica das RNRs. A princípio há um caminho alternativo, diferente do caminho das máquinas de estados, mas que também soluciona com sucesso o problema para o qual a RNR foi treinada. Desta forma alguns autores adicionaram aos seus algoritmos de extração de regras uma etapa de minimização dos estados obtidos [Giles et al. 1992].

Outra técnica de quantização foi proposta por Giles *et al* [Giles et al. 1991] e aprimorada por Omlin & Giles [Omlin and Giles 1996]. Nesta técnica, o espaço de estados é particionado em hipercubos idênticos, representando os macroestados. Realiza-se uma busca em largura alimentando a rede com os padrões de entrada até que nenhum novo hipercubo seja visitado. As transições entre os macroestados são a base da máquina extraída. Por fim também há a necessidade de simplificar o autômato extraído. Esta técnica apresenta dois problemas decorrentes do fato de analisar todos os possíveis estados, e para estes, todas as possíveis entradas da rede, que é a complexidade exponencial e o risco de obter uma representação da rede mais complexa do que a real, uma vez que a rede pode não trabalhar em todo o domínio possível [Jacobsson and Ziemke 2003].

Zeng *et al* [Zeng et al. 1993] propuseram uma abordagem diferente para os macroestados, utilizando o algoritmo *k-means* para clusterizar os microestados. Os centróides dos clusters são então utilizados como a base da busca em largura. Com esta abordagem os macroestados terão número e tamanho de acordo com o domínio da aplicação. No entanto, a determinação dos k clusters é um parâmetro difícil de ser definido, e que, se mal definido, pode ocasionar erros já na primeira etapa da extração: a discretização do espaço de estados. Uma alternativa para a determinação correta do número de clusters foi formá-los durante o treinamento da RNR [Frasconi et al. 1996].

Outra abordagem, baseada na amostragem de RNR, foi proposta por Watrous & Kuhn [Watrous and Kuhn 1992]. Esta proposta consiste em armazenar as interações da RNR com os dados, de forma a determinar um domínio de atuação para a rede, o qual funcinaria como uma heurística. Tal heurística faria com que

fossem analisados somente os estados relevantes. Com esta abordagem o conjunto de regras extraído se torna menor, evitando informações irrelevantes. No entanto um problema foi apresentado, a ocorrência de inconsistências nas transições, o que iria contra a formalismo das máquinas de estados. Estas inconsistências serviram como motivação para a utilização de outra forma de descrição do conhecimento extraído, na forma de máquinas estocásticas.

Tiño & Vojtek [Tino and Vojtek 1998] propuseram uma metodologia para extração de máquinas estocásticas. Apesar de através da técnica de amostragem ocorrerem transições inconsistentes, tais transições podem apresentar um padrão de ocorrência. Através da contagem destas transições podem ser calculadas suas propabilidades de ocorrência, as quais podem ser interpretadas como as probabilidades de transição de uma máquina estocástica.

A representação de conhecimento, extraído de RNRs, através de cadeias de Markov, foi abordada por Pechmann & Cechin [Pechmann and Cechin 2005, Pechmann and Cechin 2004, Pechmann 2004] que utilizaram como métodos de clusterização o *k-means* e também a clusterização difusa. Nestes trabalhos observou-se que este tipo de representação demonstrou bons resultados para os sistemas dinâmicos analisados. As cadeias de Markov extraídas apresentaram comportamento idêntico ao dos sistemas utilizados para o treinamento das RNRs.

Enfim, vários autores vêm estudando a extração de conhecimento de RNRs em diferentes domínios, apresentando resultados promissores. No entanto, não foi encontrado nenhum trabalho utilizando esta abordagem para análise de microarranjos, o que é o objetivo deste trabalho.

4.7 Discussão e encerramento do Capítulo

Este Capítulo buscou apresentar ao leitor as principais técnicas computacionais empregadas em análise de dados de microarranjos. Pode-se observar a grande quantidade de técnicas empregadas bem como a falta de uma metodologia padrão para as análises.

Muitos trabalhos buscam comparar diversas técnicas, mas, inclusive estes, encontram dificuldades em apontar uma das alternativas como sendo a mais adequada. O desempenho das técnicas aplicadas depende diretamente dos dados coletados, podendo, por exemplo, uma determinada técnica ter melhor ou pior desempenho de acordo com os Microarranjos analisados.

Desta forma, buscou-se na bibliografia, técnicas e métodos de análise que melhor se adaptem às realidades deste trabalho. Todos os tipos de análises citados tem sua importância, no entanto, cada um com a sua abordagem.

A seleção de características é vital para o processo, no entanto, para este experimento, não se necessita de um método tão exato e determinístico como os propostos por Chuang *et al* [Chuang et al. 2003] e Holter *et al* [Holter et al. 2000]. Tais métodos possuem parâmetros difícies de serem determinados, que necessitariam de um estudo a parte sobre sua influência na análise, o que acarretaria em um período de tempo impraticável para sua aplicação em uma metodologia como esta.

Para esta metodologia seria ideal reunir uma ténica de seleção que una características de visualização, com a possibilidade de agrupamento dos dados. Com este fim, uma técnica utilizada com sucesso tem sido os Mapas Auto-organizáveis [Toronen et al. 1999, Slonim 2002, Tamayo 1999, Oja et al. 2002, Nikkila et al. 2002]. Com esta ferramenta é possível visualizar os dados de forma a determinar as características mais representativas, e após isto, determinar os grupos que apresentam tais características de forma clara e sem a necessidade de conhecimento prévio dos dados analisados.

Quanto às técnicas de extração de conhecimento aplicadas a dados de Microarranjos, ainda não há uma técnica eficiente para tal análise. Butte *et al* [Butte and Kohane 2000, Butte 2002] propõe métodos de extração baseado na análise de correlação entre os dados. Tal método é matematicamente eficiente, no entanto, quando os dados de interesse possuem milhares de variávies, o resultado pode se tornar extremamente difícil de ser analisado.

Outra proposta eficiente, mas que não se adapta a grandes volumes de informação, são os modelos contínuos. Existem propostas de modelagem dos dados de Microarranjos através de equações lineares (ou não lineares, de acordo com a complexidade) [Styczynski and Stephanopoulos 2005], e até mesmo utilizando o modelo *S-system* [Sakamoto and Iba 2001, Noman and Iba 2005], que é aplicado para descrição de redes bioquímicas. No entanto estas técnicas não suportam um grande volume de dados, e a determinação dos parâmetros destes modelos/equações é de alto custo computacional.

Quanto aos modelos estocásticos utilizados, grande parte dos trabalhos utilizam Redes Bayesianas com uma abordagem temporal, a qual deram o nome de Redes Bayesianas Dinâmicas (DBN). Este modelo adiciona a representação temporal das Cadeias de Markov às Redes Bayesianas [Murphy and Mian 1999, Ong et al. 2002, Kim et al. 2003, Perrin et al. 2003]. No entanto as DBNs apresentam resultados pouco confiáveis. Zhou *et al* [Zhou et al. 2004] utilizou, em conjunto com as DBNs, o método MCMC (Monte Carlo Markov Chain), conseguindo melhores resultados, mas, ainda não satisfatórios.

Assim, o processo de Extração de Conhecimento a partir de dados de Microarranjos ainda não possui um consenso sobre qual a metodologia a ser utilizada. Este trabalho propõe aplicar a metodologia de Extração de Conhecimento a partir de RNRs proposta por Pechmann & Cechin [Pechmann and Cechin 2005, Pechmann and Cechin 2004, Pechmann 2004] na forma de Cadeias de Markov, visando inferir as relação presentes nos dados de Microarranjos.

Como visto, as RNRs são teoricamente equivalentes a máquinas de Turing, podendo modelar qualquer tipo de informação [Siegelmann and Sontag 1992]. Além disso, elas apresentam boa habilidade em modelar dados temporais [Pechmann and Cechin 2005, Pechmann and Cechin 2004, Pechmann 2004]. Unindo estas características à possibilidade de se extrair o conhecimento destas redes na forma de uma Cadeia de Markov [Tino and Vojtek 1998, Pechmann 2004] acredita-se ter uma metodologia adequada à análise de dados de Microarranjos. A metodologia proposta será apresentada, em detalhes, no próximo capítulo.

Capítulo 5

Metodologia

Neste trabalho é proposta uma metodologia para encontrar as relações de influência entre os genes, a partir de dados brutos de Microarranjos. Nesta metodologia estão envolvidas diferentes técnicas computacionais a fim de extrair o conhecimento desejado.

Em poucas palavras pode-se descrever esta metodologia (Figura 5.1) nas seguintes etapas:

- **Seleção dos dados:** selecionar os dados pertinentes à análise, dentre as diversas bases de dados existentes.
- **Determinação dos padrões:** determinação dos padrões presentes nos dados, bem como o pré-tratamento necessário.
- **Composição da base de treinamento:** a partir dos principais padrões determinados, e também dos dados brutos, compor a base de treinamento para a RNR.
- **Treinamento de RNR:** determinação dos parâmetros de treinamento da RNR com a finalidade de aproximar as séries temporais oriundas dos Microarranjos.
- **Extração da Cadeia de Markov:** extração da Cadeia de Markov, a partir da RNR treinada, e da base de treinamento.

As próximas seções descrevem esta metodologia de forma detalhada, explicando cada uma das etapas.

5.1 Determinação da Base de Dados

Os dados manipulados neste trabalho são matrizes de expressão gênica, obtidas através de análises de Microarranjos. Nestas matrizes, cada linha corresponde ao nível de expressão gênica de um determinado gene, em diversas amostras no tempo, formando as *séries temporais*. Cada coluna corresponde ao nível de expressão gênica

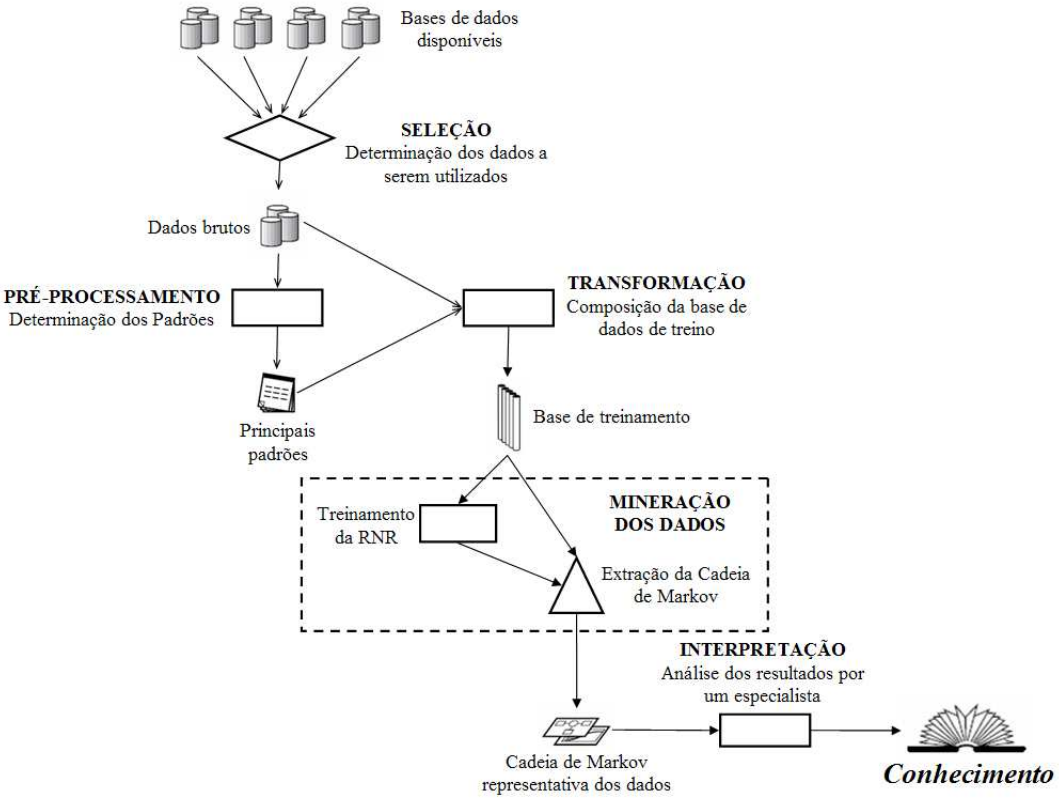


FIGURA 5.1 – *Visão global da metodologia.*

de todos os genes em um mesmo instante, a qual nos referimos como *amostra temporal*. A Figura 5.2 mostra em detalhes a organização da matriz de expressão gênica.

Atualmente, com a popularização da técnica de Microarranjos, o número de dados oriundos deste tipo de análise cresce consideravelmente. Há diversas bases de dados disponíveis na Internet. Assim, é preciso decidir por uma destas e, para isto, alguns fatores devem ser levados em consideração. Entre eles, o tamanho das séries temporais e os trabalhos relacionados.

O tamanho da série temporal deve ser o maior possível para otimizar o aprendizado da RNR. Entretanto, grande parte das bases de dados de Microarranjos possuem séries curtas, apresentando em torno de 20 amostras temporais.

Séries temporais curtas, para este tipo de análise, são desaconselháveis, pois estas séries não possuem muitas informações em seus dados, prejudicando, desta forma, o processo de extração do conhecimento.

Outro fator a ser levado em consideração, são os trabalhos utilizando os mesmos dados. O fato de outros autores estarem trabalhando com a mesma base, significa que há uma maior quantidade de informações a respeito do conjunto de genes em questão permite também a comparação com resultados já obtidos, facilitando a validação da metodologia proposta.

Após a determinação da base de dados a ser analisada, é necessário um pré-tratamento, principalmente devido ao grande volume de informações presentes nos

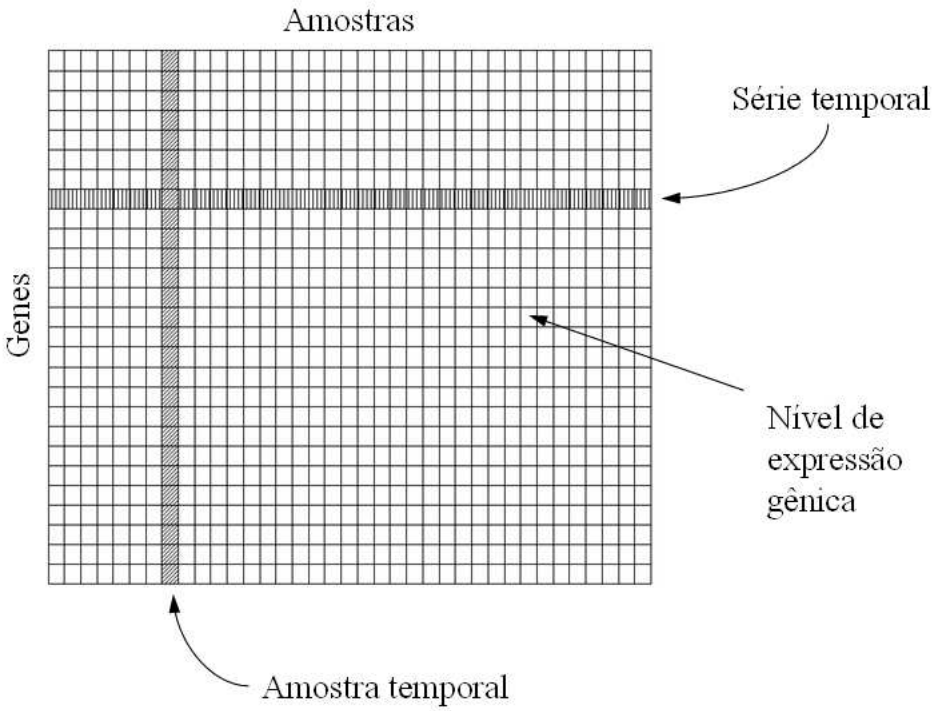


FIGURA 5.2 – *Matriz de expressão gênica.*

Microarranjos. Esta etapa é descrita a seguir, na próxima seção.

5.2 Pré-tratamento dos dados

Para a extração da Cadeia de Markov dos dados de Microarranjos, primeiramente é necessário o treinamento de uma RNR com a matriz de expressão gênica selecionada. Neste ponto, encontram-se duas características conflitantes: o fato das RNRs não comportarem muitos neurônios na camada de entrada (parâmetros de entrada) [Haykin 2001], e o grande volume de informações, típico dos microarranjos [Chuang et al. 2004, Causton et al. 2003]. Para resolver este conflito, realiza-se um pré-processamento sobre a matriz de expressão gênica, a fim de reduzir o número de parâmetros de entrada da RNR.

Para atingir tal objetivo, várias técnicas são aplicadas com sucesso [Butte 2002], entre elas a Clusterização Hierárquica [Slonim 2002], *Nearest Neighbours* [Golub et al. 1999] e Mapas Auto-organizáveis (SOM) [Tamayo 1999]. Deseja-se, com este pré-tratamento, selecionar os padrões mais representativos da base de dados. Entretanto, tais padrões não são conhecidos *a priori*. Assim, os Mapas Auto-organizáveis foram escolhidos por proporcionarem uma boa visualização dos dados, facilitando a tarefa de seleção de padrões.

5.2.1 Determinação dos Padrões de Dados

A grande quantidade de informações é uma característica presente nas medições do nível de expressão gênica. Um experimento de Microarranjo possui milhares de séries temporais, cada uma correspondendo ao nível de expressão de um gene. Desta forma, trabalhar com todos estes dados torna-se impraticável. Dentre os diversos problemas em lidar com tamanha quantidade de séries de expressão gênica, podemos destacar dois. Primeiro, estes dados serão utilizados para treinar uma RNR, sendo que cada série correspondendo a um parâmetro de entrada. Isto levaria a uma RNR com um número muito elevado de neurônios, o que não geraria bons resultados, ainda mais que o objetivo do trabalho é encontrar as relações existentes entre os dados. Em segundo lugar, analisar tal quantidade de níveis de expressão gênica poderia inviabilizar a interpretação dos resultados e a validação da metodologia.

Para resolver este problema reduz-se a quantidade de expressões gênicas a serem estudadas. A matriz de expressão gênica deixa de ser utilizada em sua totalidade e passa a ser representada por algumas séries temporais, no caso, as séries representadas pelos principais padrões presentes na matriz de expresssão gênica.

Para determinar os principais padrões utiliza-se os SOMs, treinados com toda a matriz de expressão gênica. Os SOMs têm como propriedade agrupar dados semelhantes, e não fornecem uma resposta direta ao problema de determinação de padrões. Para este fim, faz-se uma análise das freqüências de ativação de cada neurônio mediante a apresentação de todo o conjunto de dados. Cada neurônio é ativado somente quando considerado, pela rede, como sendo o representante mais próximo do dado de entrada. Considera-se, assim, o número de vezes que o neurônio foi ativado como sendo o número de séries temporais representadas pelo mesmo. Desta forma, os neurônios que apresentarem um maior número de ativações são considerados os principais padrões presentes na matriz de expressão gênica, e os pesos do neurônio como sendo o protótipo do padrão.

Para determinar o número de padrões, define-se um limiar que indica quantas séries temporais um neurônio deve representar para poder ser considerado um padrão. A Figura 5.3 exemplifica este processo apresentando o histograma das ativações dos neurônios onde é definido, como limiar, 35 séries para cada padrão. Deve-se atentar para não definir um limiar elevado, para não excluir padrões importantes presentes nos dados, bem como, limiares baixos, para evitar trabalhar com o ruído presente nos dados. No presente trabalho obteve-se 11 padrões, e as séries representadas por estes padrões serão utilizadas para construir a base de dados de treinamento das RNRs.

Através da representação dos dados da matriz de expressão gênica por seus principais padrões, ocorre uma redução no volume de informação manipulada. Isto facilita o trabalho, tanto em termos computacionais, como em termos de validação biológica. Os dados originais são apresentados na Figura 5.4(a) e, os padrões obtidos, na Figura 5.4(b).

Contudo, somente determinar os padrões não é suficiente. Após concluída esta etapa, deve-se utilizar estes padrões na construção da base de treinamento para a

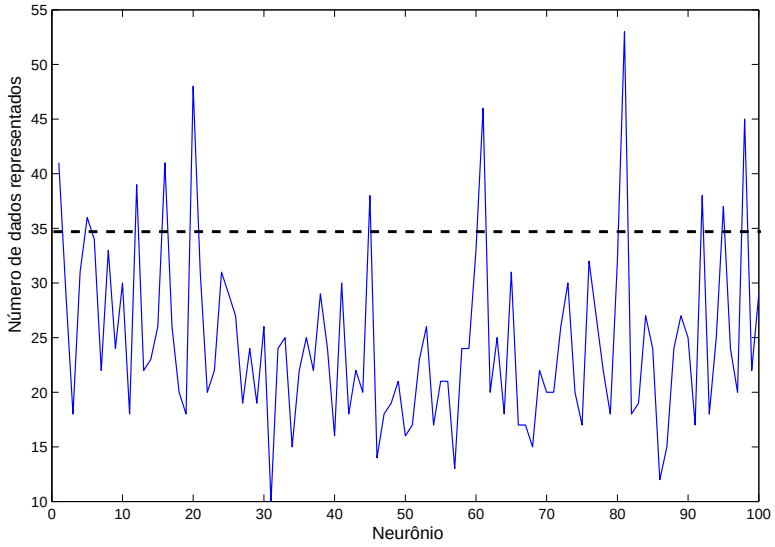
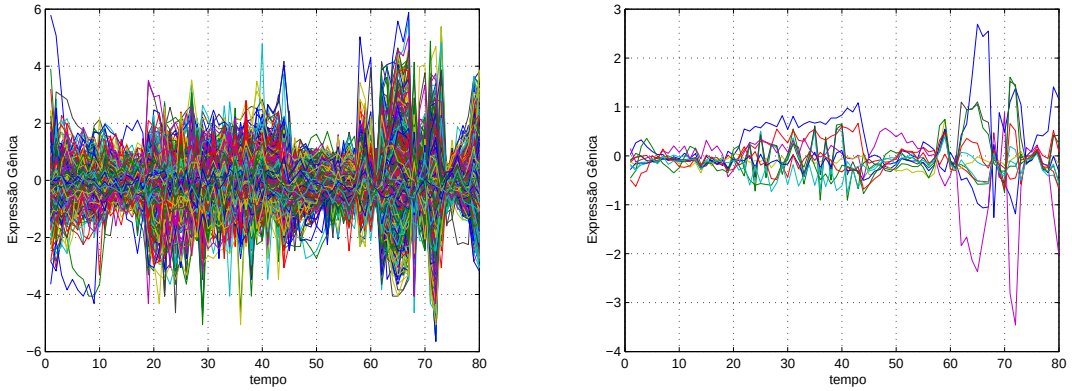


FIGURA 5.3 – *Quantidade de dados representado por cada neurônio da SOM.*



(a) Dados de microarranjos em sua forma original.

(b) Protótipo dos principais padrões determinados a partir dos dados originais.

FIGURA 5.4 – *Dados de Microarranjos.*

RNR. A explicação de como esta base é construída é apresentada na próxima seção.

5.2.2 Construção da Base de Treinamento

Uma vez determinados os padrões de dados que serão utilizados, o próximo passo é construir a base de treinamento para a RNR. Cada padrão corresponde a um dos parâmetros de entrada da Rede Neural, contudo, o padrão informado pelos SOMs não é um dado verdadeiro, e sim, um protótipo representando um valor médio de um conjunto de dados semelhantes.

A base de dados de treinamento poderia ser composta apenas pelos protótipos dos padrões, mas os resultados obtidos deste modo não considerariam diversas características presentes nos dados. Como cada padrão determinado representa, no mínimo, 36 séries temporais, em cada uma destas séries pode haver características que não ficam evidentes no protótipo do padrão, e, conseqüentemente, não seriam informadas a RNR.

Tendo isto em vista, a base de dados foi construída a partir de séries temporais originais de microarranjos. Cada parâmetro de entrada da rede neural corresponde a um dos padrões determinados. Considerando que o padrão com menor número de representantes possui K séries temporais, então, para cada parâmetro de entrada foram selecionadas K séries de cada padrão. Com isto, cada neurônio da RNR tem como base de treinamento K séries temporais, correspondentes a um mesmo padrão, dispostas de maneira sequencial. Como, cada série temporal é composta por J amostras, a matriz com os dados de treinamento possui dimensões P x $[K*(J-1)]$, onde P corresponde aos padrões (número de neurônios) e $K*(J-1)$ corresponde às séries temporais concatenadas. Para cada série, trabalha-se apenas com $J-1$ amostras por se ter removida a última amostra das séries utilizadas como entrada da RNR, e a primeira amostra das séries utilizadas para a saída. Tal procedimento foi realizado para evitar ligações entre amostras de séries diferentes.

Após estas etapas, os dados estão em uma forma adequada para o treinamento da RNR, descrito a seguir.

5.3 Treinamento da RNR

A RNR deve ser treinada de modo a aprender as principais características dos dados e conseguir reproduzir, mesmo que em poucos passos, as séries temporais apresentadas. Para isto se utilizam redes baseadas na topologia de Jordan, com uma camada oculta.

Para o número de neurônios na camada oculta, não se determinou a princípio um número de neurônios a ser utilizado. Para cada experimento, redes com diferentes números de neurônios ocultos foram analisadas com o intuito de encontrar a rede com melhor resultado. Este melhor resultado significando baixo erro de aprendizagem e boa capacidade de predição com o menor número possível de neurônios.

Conseguir minimizar o número de neurônios na camada oculta é necessário para facilitar a tarefa de extração da Cadeia de Markov.

Uma vez que as RNRs estejam treinadas com os dados de expressão gênica, o próximo passo consiste em extrair a Cadeia de Markov com os estados representativos do comportamento dos dados, e determinar as interações existentes em cada um dos estados.

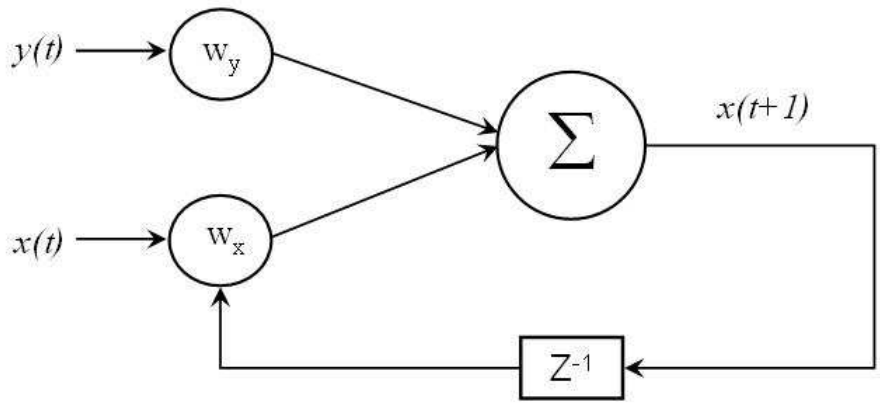


FIGURA 5.5 – *RNA Recorrente Linear.*

5.4 Extração de Conhecimento de RNRs

Uma vez que se disponha de um conjunto de dados, as RNRs podem ser utilizadas como aproximadores universais de funções, comportando-se da mesma forma que o processo gerador dos dados. Assim, a rede passa a representar um modelo limitado do processo, que armazena após o treinamento, as características dos dados [Pechmann 2004].

5.4.1 Introdução

Uma vez que RNRs podem ser utilizadas para aproximação de funções, pode-se dizer que a mesma implementa a função que rege o comportamento dos dados. No caso desta aplicação, lida-se com séries temporais e, considerando um exemplo bem simples, de uma rede recorrente linear, pode-se dizer que a rede reproduz a equação de recorrência que gera os dados de treinamento.

Na Figura 5.5 pode ser visto um sistema linear recorrente utilizado para aproximação de uma série de dados. Neste exemplo há duas entradas $y(t)$ e $x(t)$, sendo esta última, recorrente e correspondendo ao operador de atraso unitário Z^{-1}. Este operador de atraso unitário é a saída da rede neural no tempo anterior $t-1$. Há também os pesos w_y e w_x que atuam como coeficientes das entradas $y(t)$ e $x(t)$, respectivamente. O processamento, neste caso, é feito pelo neurônio linear de saída, que realiza simplesmente um somatório do produto das entradas pelos pesos a elas correspondentes.

Considerando o mecanismo de funcionamento da rede, expressa-se matematicamente o processamento feito pela rede, sobre os dados, através da seguinte equação de recorrência:

$$x\left(t+1\right) = w_y * y(t) + w_x * x(t) \tag{5.1}$$

A rede neural modelou a série analisada. A partir da rede pode-se dizer que

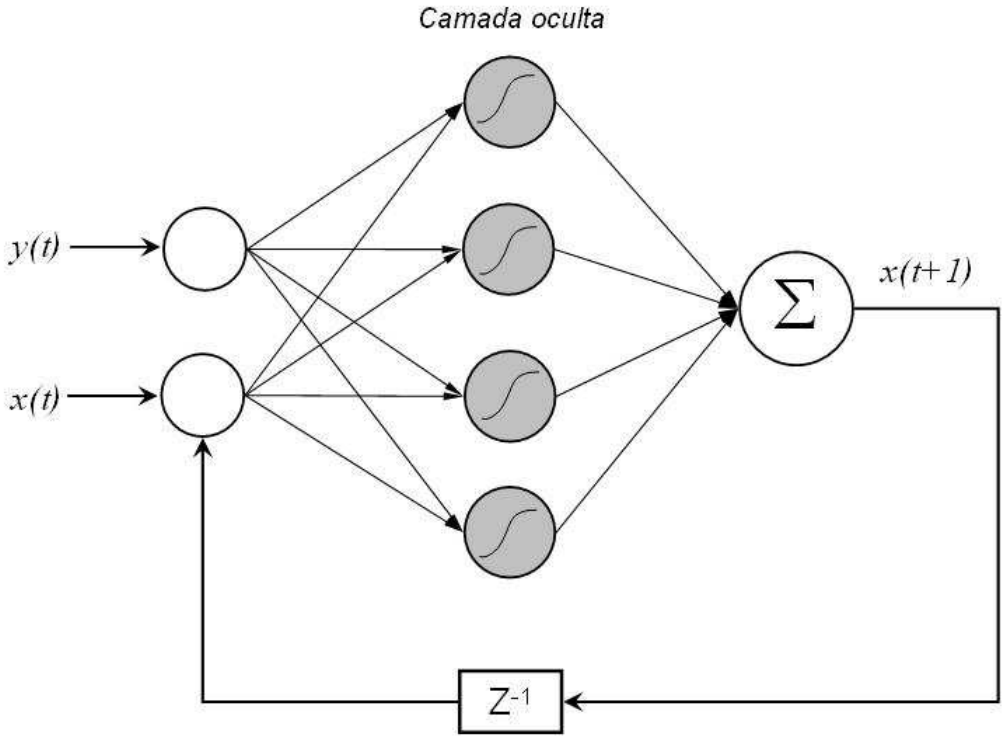


FIGURA 5.6 – *Modelo de uma RNA Recorrente Não-Linear com uma camada oculta.*

a saída $x(t+1)$ sofre uma influência w_y da entrada $y(t)$, e uma influência w_x da entrada $x(t)$. Estas influências são fornecidas diretamente pelos pesos da rede, os quais são ajustados no momento do treinamento, e através da Equação 5.1 observa-se matematicamente o seu significado.

Como este trabalho manipula dados de microarranjos, os quais possuem uma marcante característica não-linear, é necessária uma modelagem que consiga reproduzir tais características presentes. O que não é possível com o exemplo de aproximação linear mostrado acima. Para tal, é sugerida a utilização de uma Rede Neural Recorrente, com uma camada oculta que possua neurônios com função sigmóide (Figura 5.6).

Utilizando, como modelo para aproximação, uma RNR com neurônios não-lineares na camada oculta, consegue-se uma aproximação eficaz das séries temporais em questão, entretando, a extração das equações de recorrência deixa de ser uma tarefa simples. A camada oculta presente na RNR permite um mapeamento mais complexo entre a entrada e a saída dos dados, e, com isso, cresce também a complexidade das equações de recorrência obtidas.

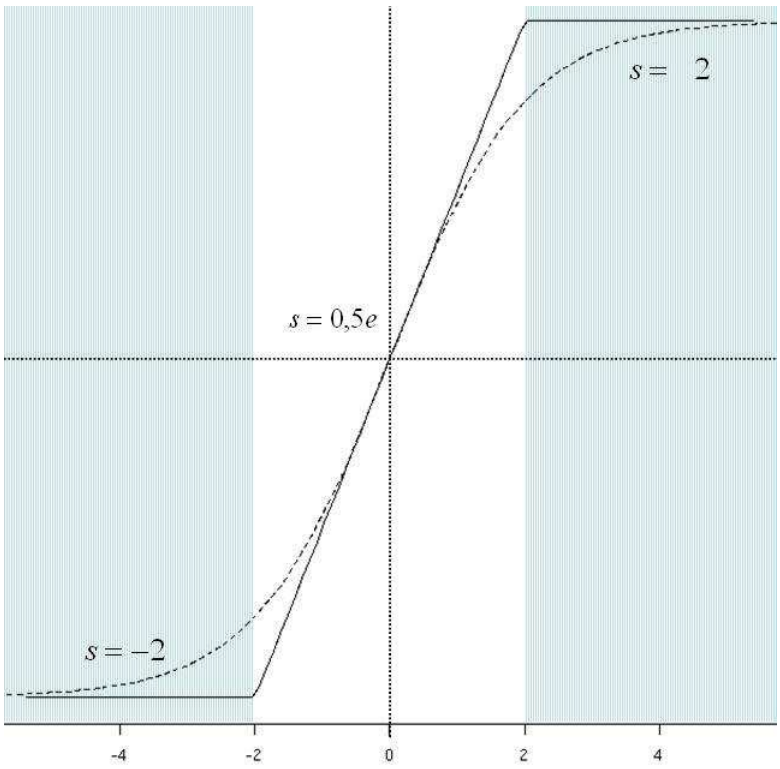


FIGURA 5.7 – *Aproximação da função sigmóide do neurônio por três funções lineares.*

5.4.2 Linearização da função Sigmóide e Equações de Recorrência

O mapeamento realizado pelas RNAs entre a entrada e a saída é feio por meio das funções de cada neurônio. Se tratando de neurônios não lineares, diversas funções de ativação podem ser utilizadas. Neste trabalho optou-se por trabalhar com funções sigmóides.

Para a construção de uma Cadeia de Markov é necessário extrair estados a partir da RNR. Estes estados são extraídos de acordo com a divisão do espaço de trabalho da rede. Este espaço é determinado através da análise da região de ativação dos neurônios da camada oculta. Estudos realizados comprovam que as ativações dos neurônios concentram-se basicamente em 3 regiões: na região central da sigmóide, e nas regiões iniciais e finais, conforme pode ser visto na Figura 5.7 [Cechin 1997]. Assim, cada uma destas regiões pode ser aproximada, por uma equação linear, com um erro de aproximação pequeno ao ponto de poder se desconsiderado.

Citando, novamente, a Figura 5.7 como exemplo, a determinação da região de trabalho do neurônio pode ser feita de diversas maneiras, tanto através dos valores de entrada do neurônio, dos valores de ativação, ou, até mesmo, através de funções de pertinência. Para exemplificar vamos considerar os valores de entrada e do neurônio. Neste caso divide-se o espaço de entrada em 3:

- para $e < -2$: o neurônio está trabalhando na primeira região, e sua saída pode ser aproximada por $s = -2$, ou seja, uma vez que o neurônio encontra-se nesta

região, seu comportamento passa a não depender dos valores de entrada;

- para $e \geq -2$ & $e \leq 2$: o neurônio encontra-se na região central, neste caso aproxima-se sua ativação por $s = 0,5e$, logo, diferentemente da primeira região, esta região sofre influência direta dos valores de entrada;
- e por fim $e > 2$: significa que o neurônio estará trabalhando na terceira região. Esta, assim como a primeira região, pode ser aproximada por uma constante, neste caso $s = 2$.

Desta forma, cada neurônio divide o espaço de dados em 3 regiões, de acordo com a função sigmóide. Considerando uma rede com vários neurônios na camada oculta, as regiões aumentam na ordem de 3^n, onde n é o número de neurônios ocultos. As combinações entre as regiões definidas em cada neurônio da rede representam possíveis estados da Cadeia de Markov a ser extraída. Contudo, como cada estado da Cadeia de Markov é definido pela combinação das regiões **ativadas** nos neurônios, nem todas as combinações precisam ocorrer, sendo assim, podem existir estados possíveis que não chegam a ser ocupados pelo conjunto de dados em questão, não tendo necessidade de fazer parte da Cadeia de Markov.

As Figuras 5.8 e 5.9 apresentam um exemplo de uma RNR com 1 neurônio na camada oculta e a respectiva divisão do espaço de trabalho deste neurônio a fim de melhor exemplificar a metodologia. Como pode ser visto, há 3 regiões de trabalho, as quais correspondem a um intervalo de dados de entrada. Quando as entradas encontram-se nas regiões escuras da Figura 5.9, o comportamento da saída pode ser aproximado por uma constante (ver Figura 5.7). Nestes casos, se a rede se encontra na região mais à esquerda, podemos aproximar os valores de saída por:

$$x(t+1) = -2 * w \tag{5.2}$$

No caso da rede encontrar-se na região mais à direita, a aproximação pode ser feita por:

$$x(t+1) = 2 * w \tag{5.3}$$

Nota-se que em ambos os casos os valores de entrada da rede não interferem nos valores de saída. Quando a rede encontra-se trabalhando na região central, isto não ocorre, pois, como foi dito anteriormente, e explicitado na Figura 5.7, a aproximação da saída s é feita por $s = 0,5e$, onde e são os valores de entrada do neurônio. Desta forma, para a rede da Figura 5.8 quando ela encontra-se trabalhando na região central, a saída é aproximada por:

$$x(t+1) = 0,5w\left[(w_y * y(t)) + (w_x * x(t)) + (w_b * b)\right] \tag{5.4}$$

5.4.3 Interpretação das Equações de Recorrência

Neste trabalho lida-se com dados de Microarranjos e busca-se a interação existente entre eles. Desta forma, as equações de recorrência, determinadas através

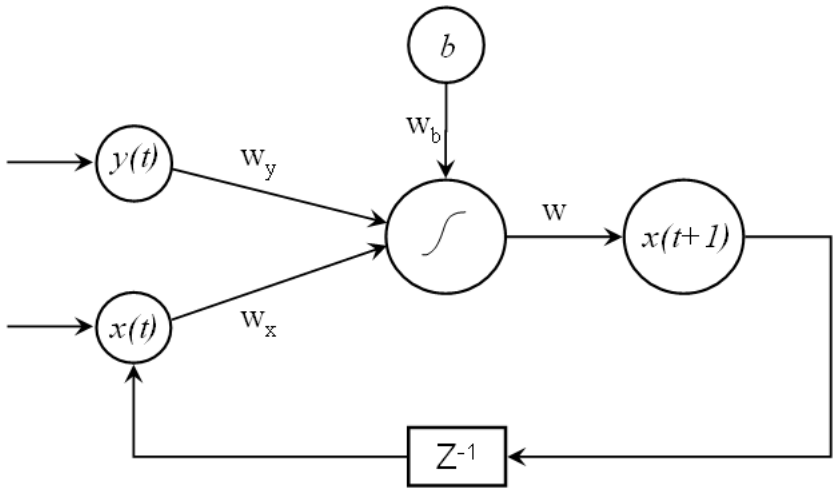


FIGURA 5.8 – *Modelo Neural Recorrente simplificado.* $y(t)$ *e* $x(t)$ *são as entradas e* b*, o* bias*;* w_y *e* w_x *e* w_b *são os respectivos pesos das entradas do neurônio sigmóide e* w *o peso atribuído a sua saída;* $x(t+1)$ *a saída da rede e* Z^{-1} *o operador de atraso unitário.*

das RNRs, podem nos fornecer as informações necessárias. Como interpretar estas informações, será apresentado nesta subseção.

Considerando o exemplo simplificado da Figura 5.8, aplicado a dados de Microarranjos, define-se as entradas $y(t)$ e $x(t)$ da RNR como sendo o nível de expressão gênica de um grupo de genes, cujo comportamento se assemelha bastante, e b, como sendo um *bias*. Neste caso temos $x(t+1)$ como saída da rede. Esta saída é a expressão gênica do grupo de genes x no tempo seguinte ao da entrada.

Assim, uma vez definidos os parâmetros da rede, pode-se analisar a Equação 5.4. Esta equação, obtida através da análise da RNR, nos fornece o valor no tempo $t+1$ de expressão gênica do grupo x, quando apresentadas para a rede as expressões gênicas dos grupos x e y. Ou seja, determina-se de que forma os parâmetros de entrada x e y influenciam no próximo valor de x.

Interpretando biologicamente esta equação de recorrência, pode-se dizer que o valor da expressão gênica $x(t+1)$ do grupo de genes x, sofre as seguintes influências:

- w_x de seu próprio nível de expressão gênica ($x(t)$);
- w_y do nível de expressão gênica do grupo y ($y(t)$);
- e $b * w_b$ atribuído a outos fatores externos aos dados apresentados.

As informações obtidas destas equações de recorrência extraídas das RNRs possuem características muito próximas das fornecidas por uma Rede Regulatória, o que fica ainda mais evidente quando analisamos estas equações em uma forma gráfica, como apresentado na Figura 5.10.

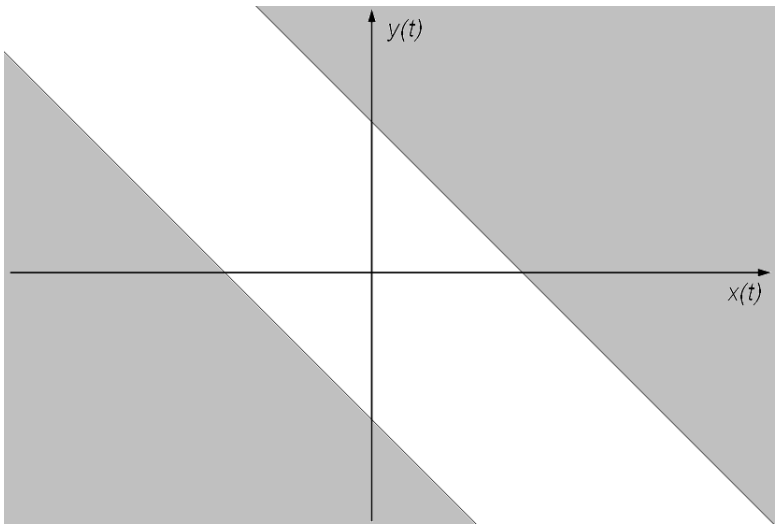


FIGURA 5.9 – *Divisão do espaço de trabalho do Neurônio sigmóide, considerando os valores de* $w_y = w_x = 1$ *e* $w_b = 0$*, da rede apresentada na Figura 5.8, em função de seus parâmetros de entrada* $y(t)$ *e* $x(t)$*.*

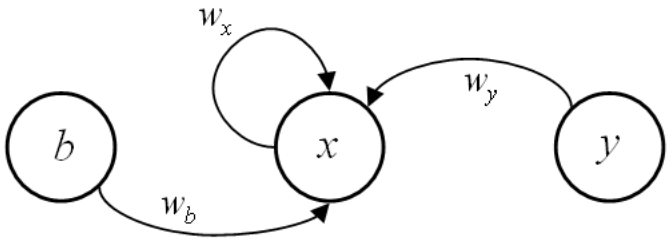


FIGURA 5.10 – *Representação gráfica da Equação 5.4.*

Uma vez já apresentados os princípios que fundamentam a metodologia, explicando e exemplificando a extração desejada sobre um modelo simples, a seção seguinte apresentará o Sistema de Extração de Regras utilizado.

5.5 O Sistema de Extração de Regras

Esta metodologia propõe a Extração do Conhecimento na forma de uma Cadeia de Markov. Para tanto é necessário, primeiramente, extrair regras, a partir da RNR, que descrevam o comportamento dos dados analisados. Para este fim, foi utilizado, como sistema extrator de regras, o FAGNIS, proposto por Cechin[Cechin 1997], mas em uma versão implementada no ambiente R[R Development Core Team 2005].

A escolha do FAGNIS, para ser utilizado como Sistema de Extração de Regras, deve-se a diversos fatores, os principais são:

- permitir a extração de regras diretamente da RNR treinada;
- garantir a equivalência de funções entre a RNR e o sistema de inferência extraído;
- apresentar uma alta compreensibilidade das regras extraídas (*i.e.* regras mais simples e em menor quantidade);
- estar implementado no ambiente R, permitindo integração com as outras ferramentas e scripts utilizados neste trabalho.

Um outro fator importante para sua escolha foi o fato de o FAGNIS resultar em um conjunto de regras que descrevem o comportamento de uma RNR em uma simples função linear, e permite a extração da influência das entradas nas saídas.

O FAGNIS realiza a partição da região de trabalho da RNR em regiões lineares de forma semelhante a regressão linear por partes, à medida que a rede vai sendo treinada. Cada região linear encontrada pelo FAGNIS corresponde a uma regra, para qual são reportados:

- *o protótipo:* que descreve o comportamento típico dos padrões pertencentes à regra;
- *a equação linear:* que determina a saída calculada para os padrões correspondentes à regra, a partir de uma análise dos pesos e ativações da rede;
- *a função de pertinência:* que indica o grau de pertinência do padrão para com a regra.

5.5.1 Definição dos Estados da Cadeia de Markov

Neste trabalho, a RNR utilizada é mais complexa do que a apresentada no exemplo. Ela possui mais neurônios de entrada e também mais neurônios na camada oculta, mas, a princípio, não se pode definir *a priori* a sua topologia, pois esta deve ser definida de acordo com os dados em questão. Assim, a partir do momento em que se trabalha com uma RNR com mais de um neurônio, o espaço de possíveis estados se multiplica, passa a ter complexidade $O(3^n)$, onde n é o número de neurônios ocultos.

Para definir os estados da Cadeia de Markov, analisa-se as ativações dos neurônios da camada oculta para todos os dados, marcando, para cada um, qual foi o estado ativado. Nem todos os possíveis estados são utilizados para representar os dados, sendo assim, os que não forem utilizados podem ser descartados da Cadeia de Markov.

O espaço de entrada da RNR fica dividido em diversos subespaços, cada um representado por um estado da Cadeia de Markov. Por sua vez, os estados da Cadeia de Markov representam o comportamento dos parâmetros de entrada da rede (grupos de expressões gênicas), mapeando as iterações presentes. Lembre-se que cada estado da Cadeia somente é válido em seu subespaço, sendo definido por um conjunto de equações de recorrência, de acordo com as regiões em que cada neurônio trabalha.

Desta forma, cada estado da Cadeia de Markov possui, associado a ele, uma Rede Regulatória válida para um conjunto dos dados de entrada. A Rede Regulatória é representada por um sistema de equações lineares definido em conseqüência da regra extraída.

5.5.2 Determinação das Transições entre os Estados

As Cadeias de Markov são definidas pelos estados e as transições entre eles. Até este ponto, as transições entre os estados não foram definidas e serão abordadas nesta seção.

A definição das transições entre os estados é feita a partir da análise do comportamento da série temporal. Apresenta-se cada um dos valores de entrada para a RNR já treinada, e, então, são analisados os valores de ativações dos neurônios da camada oculta. A combinação das regiões que estão sendo ativadas em cada neurônio indicará o estado no qual a RNR está trabalhando. Ao fazer isto, consegue-se determinar o estado da rede para cada um dos dados da série.

A Figura 5.11 representa a análise da série dos dados, neste caso, em uma rede com apenas um neurônio na camada oculta. Estão representados os 3 estados possíveis de trabalho da rede e os pontos pretos representam os dados. Os dados estão postos na figura de acordo com a região ativada no neurônio. As setas representam as transições presentes entre os dados. As setas finas representam as transições entre um mesmo estado, já as setas largas representam as transições entre os estados.

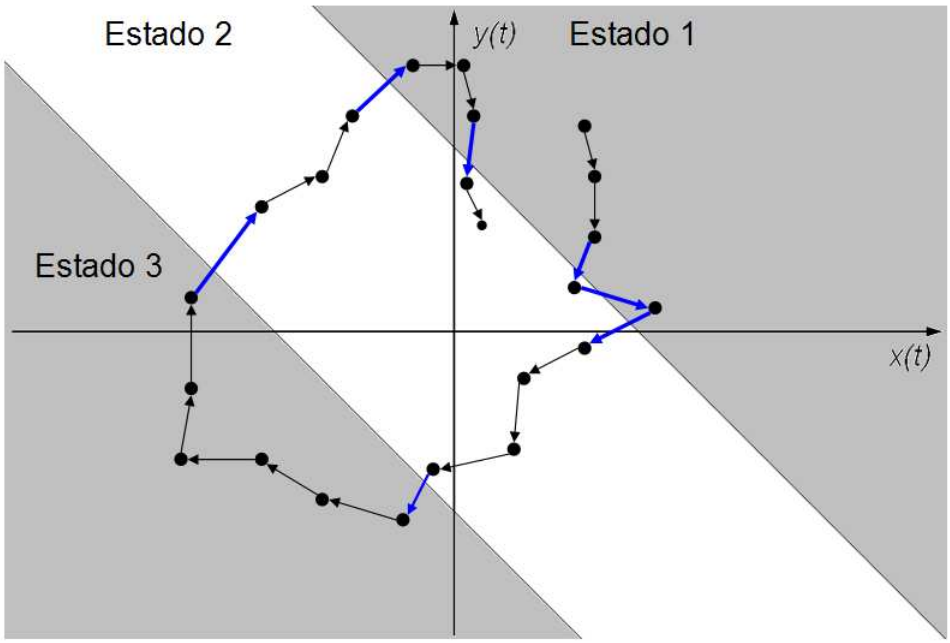


FIGURA 5.11 – *Representação da análise temporal dos dados para determinação da transição entre os estados.*

Assim, o próximo passo é calcular a matriz de contagem de transições. Nesta matriz estão todas as transições presentes nos dados. A Tabela 5.1 mostra esta

matriz para os dados da Figura 5.11.

TABELA 5.1 – Matriz de contagem de transições.

Estados	1	2	3
1	4	3	0
2	2	6	1
3	0	1	5

Na matriz de contagem, a linha representa o estado $e(n)$ e a coluna representa $e(n+1)$. A partir desta matriz faz-se um estudo estatístico das transições, calculando suas probabilidades de ocorrência. Como resultado deste estudo tem-se a Matriz de Transição de Estados, apresentada na Tabela 5.2.

TABELA 5.2 – Matriz de Transição de Estados.

Estados	1	2	3
1	$^4/_7$	$^3/_7$	0
2	$^2/_9$	$^6/_9$	$^1/_9$
3	0	$^1/_6$	$^5/_6$

Tendo os estados e as transições existentes entre eles, a Cadeia de Markov já pode ser representada de forma gráfica. A cadeia cujas transições foram determinadas nesta seção é apresentada na Figura 5.12.

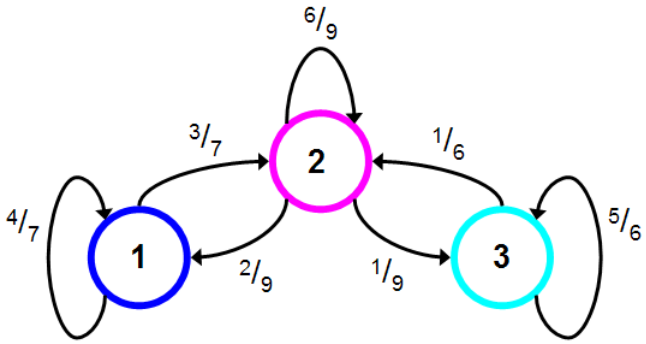


FIGURA 5.12 – *Cadeia de Markov com as transições definidas pela Matriz de Transição da Tabela 5.2.*

A seguir uma breve discussão sobre o Capítulo e o seu encerramento.

5.6 Encerramento do Capítulo

Neste capítulo, foi apresentada a metodologia proposta para determinar a interação entre genes através da análise de dados de Microarranjos.

Foram apresentadas as técnicas adotadas para pré-processar os dados, o procedimento de treinamento da RNR, o fundamento no qual está baseada a extração de conhecimento, bem como o sistema de extração de regras para a obtenção da Cadeia de Markov.

Para validar esta metodologia, foram realizados alguns experimentos que serão apresentados no capítulo seguinte.

Capítulo 6

Experimentos e Resultados

Este capítulo apresenta os quatro experimentos realizados, juntamente com uma discussão detalhada dos resultados de cada um. Cada experimento possui cinco etapas: seleção dos dados, determinação dos padrões, composição da base de treinamento, treinamento da RNR e, por fim, extração da Cadeia de Markov. As etapas de cada experimento seguem a metodologia descrita, e suas particularidades serão discutidas separadamente.

O primeiro experimento mostra um caso particular, onde os dados foram gerados artificialmente, desta forma, dispensando as etapas de seleção e determinação de padrões. Os dados foram gerados através de três equações diferenciais e o objetivo deste experimento é lidar com uma base de dados cuja relação existente entre as variáveis seja conhecida, falicitando, desta forma, a validação da metodologia. Os demais experimentos utilizam dados de Microarranjos.

O segundo experimento analisa os dados do Controle de Divisão Celular (CDC-15). A principal meta deste experimento é identificar o ciclo presente nestes dados. O comportamento cíclico destes genes é conhecido pelo meio científico [Spellman et al. 1998], assim, este deve ficar evidente na Cadeia de Markov obtida.

O terceiro experimento foi realizado com os dados do Stanford Microarray Database [Ball et al. 2005], base de dados que possui o maior número de amostras temporais entre as analisadas. Esta base possui expressões gênicas envolvidas em diversas etapas da vida celular, sendo assim, com comportamentos diferentes. Os resultados obtidos neste experimento se mostraram de difícil análise, necessitando de um especialista na área. Contudo, algumas conclusões foram possíveis a respeito das relações extraídas pela metodologia. Estas conclusões serão apresentadas, juntamente com uma breve discussão de seu significado.

Com a metodologia aplicada a uma base de dados artificiais, a duas bases conhecidas, e com os resultados já analisados, o último experimento realizado utiliza uma base de dados ainda não estudada. Esta última análise visa encontrar relações causais entre os genes codificantes de proteínas envolvidas em canais de íons, do ouvido interno de ratos, durante o desenvolvimento da audição. Os experimentos e uma discussão dos resultados também são apresentados neste capítulo.

6.1 Experimento 1 - Dados Artificiais

Este experimento surgiu frente a necessidade de uma validação formal da metodologia proposta. Este trabalho propõe uma metodologia para extração de conhecimento, no entanto, as bases de dados disponíveis possui um grande volume de informação, e pouco pode ser afirmado sobre estes dados. Desta forma, não há como comparar o conhecimento extraído para poder validar a metodologia.

Uma alternativa a este problema foi a geração dos dados a serem analisados. Desta forma toda a informação presente nos dados seria conhecida *a priori*, logo, haveria uma saída desejada para a metodologia, permitindo, assim, determinar sua validade.

6.1.1 Seleção dos Dados

Os dados a serem manipulados neste experimento devem se aproximar ao máximo dos dados dos experimentos de Microarranjos, ao mesmo tempo, não devem ser complexos ao ponto de dificultar a análise. Assim, optou-se por gerar 3 séries temporais, através do seguinte conjunto de equações recorrentes:

$$\begin{cases} g_1 = & 1.001g_1 + 0.01g_2 \\ g_2 = & -0.008g_1 + 1.001g_2 \\ g_3 = & (logistic\,(g_1) - 0.5) + g_3 \end{cases} \tag{6.1}$$

onde:

$$logistic\,(x) = \frac{1}{1 + exp\,(-x)}. \tag{6.2}$$

A partir destas equações foram geradas 1000 amostras temporais, resultando na Figura 6.1. Nos dados observa-se um comportamento constante das curvas g_1 e g_2, mas, g_3 apresenta-se de duas maneiras, com comportamento crescente e decrescente. Assim, pode-se dizer que o comportamento de g_3 representa dois estados presentes nos dados. Estes estados devem ser obtidos na Cadeia de Markov extraída, aplicando-se a metodologia.

6.1.2 Determinação dos Padrões

No caso deste experimento, em que os dados foram gerados artificialmente, não há a presença de um grande volume de informação, uma vez que somente foi gerado um volume de dados necessário à análise. Assim, esta etapa não se fez necessária para o conjunto de dados em questão.

6.1.3 Composição da Base de Treinamento

A Base de Treinamento está dividida em dois conjuntos de dados: o conjunto de entrada e o de saída. Esta composição foi feita diretamente com os dados gerados pelo conjunto de equações 6.1.

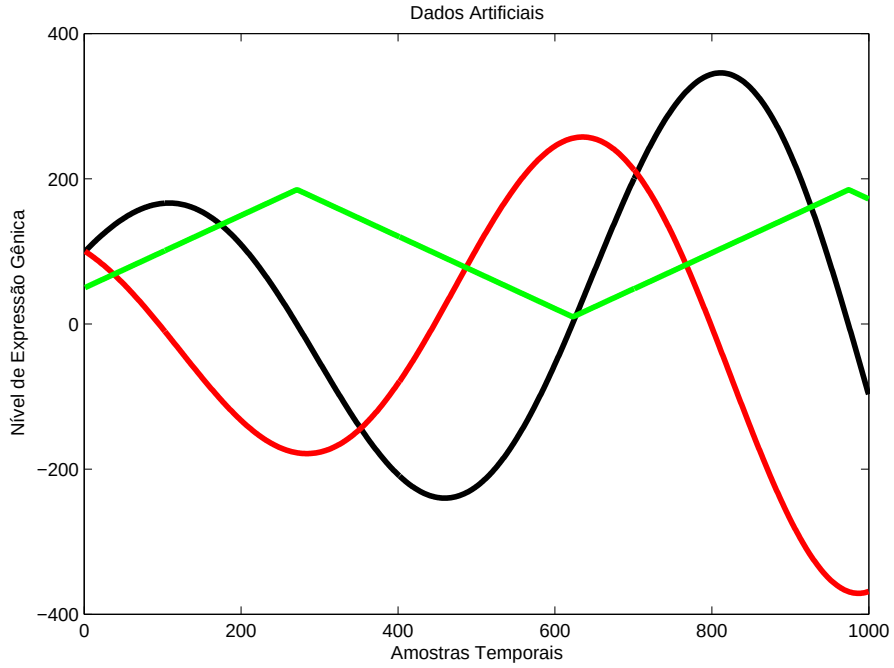


FIGURA 6.1 – Dados Artificiais: g_1 em preto, g_2 em vermelho e g_3 em verde.

No entanto, antes de separar a base de dados em conjuntos de entrada e saída, ela foi padronizada, removendo a média e dividindo pelo desvio padrão [Law and Kelton 2000]. Com a base padronizada, procede-se com a composição dos conjuntos de dados.

Os dados gerados consistem em uma série temporal com 1000 amostras. A partir destes dados foram compostos os conjuntos de entrada e saída de forma a proporcionar recorrência a rede. O conjunto de entrada é composta das amostras de 1 a 999 dos dados e o conjunto de saída das amostras de 2 a 1000.

Uma vez determinados os conjuntos de entrada e saída, a próxima etapa consiste no treinamento das RNR.

6.1.4 Treinamento da RNR

A etapa de Treinamento das RNRs não deve ser vista como uma etapa independente das demais. Esta pode ser considerada como o meio da metodologia, e está intimamente ligada tanto às etapas anteriores, como as posteriores.

As etapas anteriores determinam a número de entradas que terá a RNR. Desta forma, a determinação dos padrões precisa ser eficiente de forma a não obrigar a camada de entrada a ter um grande número de neurônios, o que poderia inviabilizar o aprendizado da rede. No caso deste experimento, devido ao pequeno número de séries temporais em questão (apenas 3), isto não foi um problema. Já as etapas posteriores são importantes na determinação do número de neurônios ocultos da RNR, os quais influenciam o número de estados obtidos para a Cadeia de Markov

extraída.

Para este experimento, devido à simplicidade dos dados em questão, foram treinadas somente 3 variações do número de neurônios da camada oculta. Foram treinadas redes com 1, 2 e 3 neurônios, tendo sido percebida pouca diferença de desempenho entre elas. Todas as redes extraíram apenas 3 estados para a Cadeia de Markov, e quanto ao erro do aprendizado, as 3 mostram resultados semelhantes. No entanto, a RNR com 1 neurônio oculto apresentou o menor erro entre as 3. Estes números encontram-se na Tabela 6.1.

TABELA 6.1 – Comparativo de desempenho das RNR treinadas com os dados artificiais.

Nro. de neurônios na camada oculta	Número possível de estados	Número de estados extraídos	Erro RMS
1	3	3	0.0002415610
2	9	3	0.0003720583
3	27	3	0.0002768718

Analisando os resultados, observa-se que o aumento no número de neurônios da camada oculta não proporciona, para estes dados, aumento da informação extraída (os quais podem ser mensurados pelo número de estados). Além disto, as RNR com 2 e 3 neurônios apresentaram erro maior que a RNR com apenas 1 neurônio. Sendo assim, a RNR com 1 neurônio mostra-se como a de melhor desempenho.

A Figura 6.2 mostra a RNR utilizada neste experimento. Esta rede segue a arquitetura de Jordan e apresenta 3 neurônios nas camadas de entrada e saída, e apenas um neurônio na camada oculta. As entradas da rede são os dados amostrados no tempo t, e as saídas, os dados amostrados no tempo $t+1$.

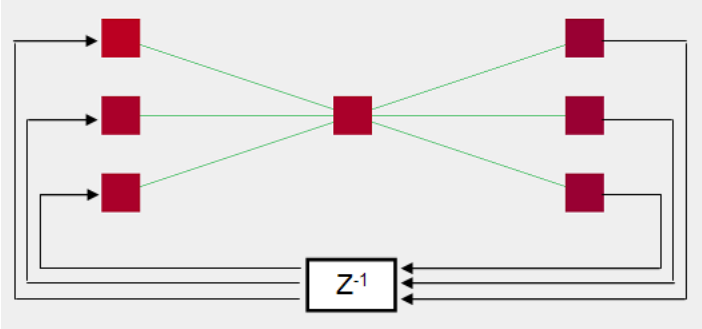


FIGURA 6.2 – *Estrutura da RNR utilizada no experimento com os dados artificiais, apresentando uma topologia* $3-1-3$*, baseada na Rede de Jordan.*

6.1.5 Extração da Cadeia de Markov

Como foi mencionado na seção anterior, 3 estados são suficientes para descrever os dados analisados, pois, mesmo aumentando o número de neurônios na camada

oculta, não ocorre um aumento no número de estados extraídos.

Como o número de estados é pequeno, não foi feito um estudo sobre uma possível redução neste número. A Figura 6.3 mostra a Cadeia de Markov extraída. Foram obtidos 3 estados, denominados de estado 1, 2 e T, e identificadas as transições existentes.

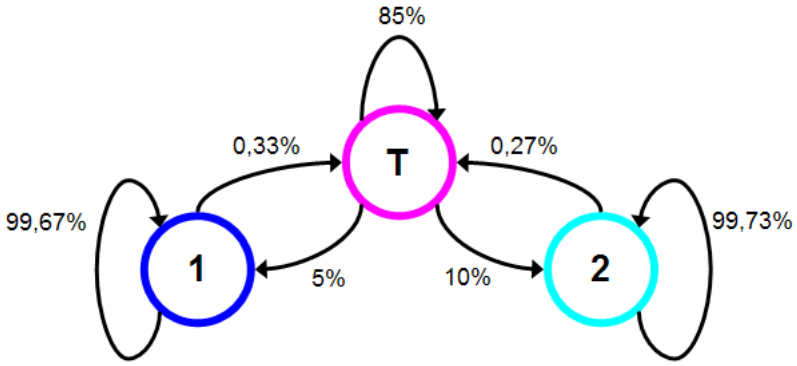


FIGURA 6.3 – Cadeia de Markov extraída do conjunto de dados artificiais.

6.1.6 Análise e Discussão dos Resultados

Uma das principais motivações para este experimento é a validação da metodologia proposta. Os dados foram gerados, e como conseqüência disto, as equações que regem seu comportamento, bem como as relações existentes, são conhecidas.

O primeiro passo é verificar, para cada estado, os dados representados por eles. A relação entre os dados e os estados pode ser vista na Figura 6.4. Através da análise desta figura observa-se que a RNR dividiu os dados em dois principais estados, representados pelo Estado 1 e pelo Estado 2, e há também um estado de transição entre estes dois, representado pelo Estado T na Cadeia de Markov.

A Figura 6.4 também explica a baixa probabilidade de transição entre os estados. É característico deste conjunto de dados permanecer muito tempo em um mesmo estado, logo, apresenta uma grande possibilidade de uma transição, de um tempo t para $t + 1$, levar ao mesmo estado. Outra consideração é quanto a ausência de transições entre os estados 1 e 2, este tipo de transição ocorre através do estado de transição T.

Enfim, as características da Cadeia de Markov extraída mostram-se representativas dos dados analisados, no entanto, também é necessário um estudo sobre as equações características dos estados para validar a metodologia. Para isto foram analisados as equações dos estados 1 e 2, por representarem quase a totalidade dos dados. Assim, como descrição para o Estado 1 temos o seguinte grupo de equações:

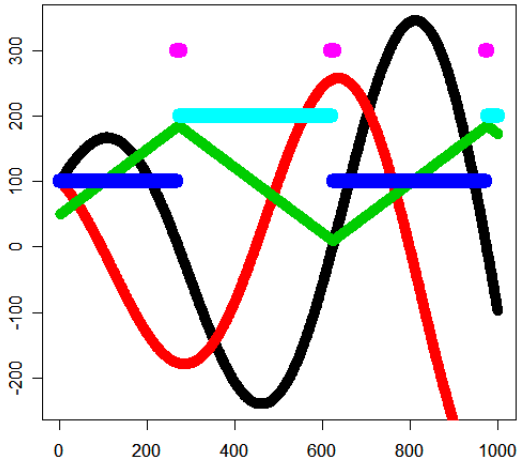


FIGURA 6.4 – Dados Artificiais com seus respectivos estados da Cadeia de Markov extraída: g_1 em preto, g_2 em vermelho e g_3 em verde, Estado 1 em azul, Estado 2 em turquesa e Estado T em magenta.

$$\begin{cases} g_1 = & 1.001\, g_1 + 0.009798\, g_2 - 2.98 \times 10^{-17}\, g_3 \\ g_2 = & -0.00816\, g_1 + 1.001\, g_2 - 4.16 \times 10^{-17}\, g_3 \\ g_3 = & -5.773 \times 10^{-10}\, g_1 - 1.39 \times 10^{-9}\, g_2 + 1.000\, g_3 \end{cases} \tag{6.3}$$

e o Estado 2 definido por:

$$\begin{cases} g_1 = & 1.001\, g_1 + 0.009798\, g_2 - 9.59 \times 10^{-19}\, g_3 \\ g_2 = & -0.00816\, g_1 + 1.001\, g_2 - 8.286 \times 10^{-17}\, g_3 \\ g_3 = & 1.289 \times 10^{-9}\, g_1 + 1.88 \times 10^{-9}\, g_2 + 1.000\, g_3 \end{cases} \tag{6.4}$$

Comparando as Equações 6.3 e 6.4, que definem os estados extraídos, com a Equação 6.1, que gerou os dados, fica visível que a metodologia consegue, a partir dos dados, determinar um conjunto de equações capaz de descrever estes dados em um determinado intervalo de tempo.

Sobre estas equações, algumas observações devem ser feitas:

- As equações que determinam g_1 e g_2 são determinadas com grande precisão, quando comparadas com a equação que gerou os dados. Nas equações extraídas aparece a influência de g_3, no entanto, esta influência é tão pequena, que pode, inclusive, ser atribuída a algum ruído proveniente da rede neural, e seu valor é tão pequeno que pode ser desprezado.

- A equação que definia g_3 era uma equação não-linear, com dois tipos de comportamento, crescente e decrescente. Estes dois tipos de comportamento foram responsáveis por dividir os dados em seus dois estados principais, um estado descrevendo os dados enquanto g_3 cresce, outro enquanto ele decresce.

- A descrição dos estados manteve-se fiel a realidade dos dados. Na série analisada g_1 e g_2 apresentam o mesmo comportamento, logo, apesar da metodologia

apresentar dois estados principais, a descrição destes dados manteve-se praticamente a mesma nos dois estados e apresentou diferenças significativas na descrição dos dados de g_3.

Enfim, este experimento permitiu comprovar que a Extração da Cadeia de Markov a partir de RNR nos permite encontrar uma descrição linear representativa do conjunto de dados analisados, inclusive com as relações existentes entre as séries analisadas.

6.2 Experimento 2 - CDC-15

Uma vez que a metodologia foi aplicada a um conjunto de dados com comportamento conhecido e que tenha sido comprovada a coerência e validade de seus resultados, este experimento visa analisar um conjunto de dados oriundo de Microarranjos. Os dados escolhidos não são complexos demais, a ponto de dificultar a análise dos resultados e também apresentam um comportamento conhecido pelo meio científico, de forma a validar os resultados obtidos.

Desta forma, os genes deste segundo experimento são os genes envolvidos no Ciclo Celular [Nurse et al. 1998]. Tais genes foram identificados por Spellman *et al* [Spellman et al. 1998] e as séries com seus níveis de expressão gênica encontram-se disponíveis no site do *Yeast Cell Cycle Analysis Project*[1]

6.2.1 Seleção dos Dados

Para este experimento foram utilizadas expressões gênicas de *Saccharomyces cerevisiae*, medidas durante o período de divisão celular. Os genes selecionados foram os identificados por Spellman *et al* [Spellman et al. 1998] como participantes do processo de Controle de Divisão Celular. Ao todo são 799 genes, cada um com 24 medições no tempo.

Quando compara-se os dados deste experimento aos dados do experimente anterior, pode se observar a grande diferença entre o número de séries analisadas. Anteriormente trabalhava-se com 3 séries (ver Seção 6.1.1), agora trabalha-se com 799. Como o número de séries (expressões gênicas) a serem analisadas é grande para ser passado diretamente a RNR, é necessário a realização de um pré-tratamento para diminuir o volume de dados a ser analisado. Este pré-tratamento será descrito na próxima seção.

6.2.2 Determinação dos Padrões

Frente ao grande volume de dados a ser analisado neste experimento, é necessário uma redução no número de variáveis envolvidas. Isto foi feito determinando os principais padrões presentes nos dados e desconsiderando os demais dados.

[1] http://cellcycle-www.stanford.edu

Para a determinação dos padrões foram utilizados os Mapas Auto-organizáveis (SOMs). Para estes dados foram analisadas 5 configurações de redes: 5×5, 8×8, 10×10, 15×15 e 20×20.

A rede 5×5 não conseguiu separar os padrões de maneira satisfatória. As redes 15×15 e 20×20 dispersaram demais os dados, determinando um grande número de padrões, com poucos dados cada um.

As redes 8×8 e 10×10 mostraram-se as melhores para estes dados. Ambas puderam determinar 9 padrões principais entre os dados, um número considerado bom para ser trabalhado. Quando comparados, os padrões determinados pelas duas redes mostraram-se muito semelhantes, aumentando a confiabilidade da análise. No entanto, a rede 8×8 apresentou um maior número de dados para cada padrão, fato esperado pois esta rede possui menor número de neurônios, logo, apresenta menor dispersão dos dados entre os neurônios. A Figura 6.5 apresenta estes principais padrões.

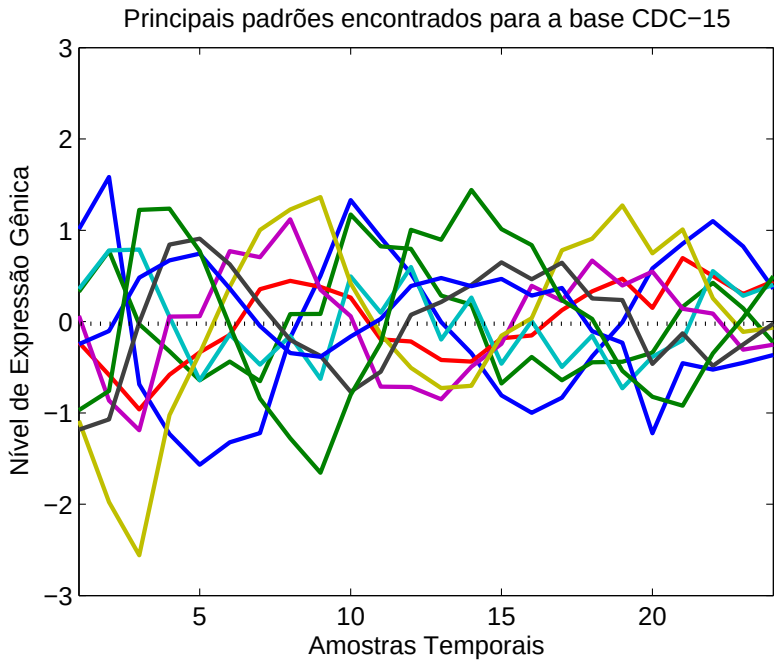


FIGURA 6.5 – *Principais padrões determinados quando analisados os dados da base CDC-15.*

Uma vez determinados os principais padrões presentes nos dados, a próxima etapa consiste da composição da base de treinamento para a RNR, que será apresentada na seção seguinte.

6.2.3 Composição da Base de Treinamento

Uma vez determinados os padrões de dados que serão utilizados, o próximo passo é construir a base de treinamentos para a RNR. Cada padrão corresponde a um dos parâmetros de entrada da Rede Neural, contudo, o padrão informado pelo

SOM não é um dado verdadeiro, e sim um protótipo que representa um valor médio de um conjunto de dados semelhantes.

A base de dados de treinamento poderia ser composta apenas dos protótipos dos padrões, mas os resultados obtidos deste modo não levariam em conta diversas características presentes nos dados. Como cada padrão determinado representa no mínimo 16 séries temporais, em cada uma destas séries pode haver características que não ficam evidentes no protótipo do padrão, e conseqüentemente não seriam informadas a RNR.

Tendo isto em vista, a base de dados foi construída a partir de séries temporais originais de Microarranjos. Cada parâmetro de entrada da rede neural corresponde a um dos padrões determinados. Considerando que o padrão com menor número de representantes possuía 16 séries temporais, optou-se por selecionar 16 séries de cada padrão, para cada parâmetro de entrada. Com isto, cada neurônio da RNR tem como base de treinamento 16 séries temporais, correspondentes a um mesmo padrão, dispostas de maneira seqüencial.

No entanto, cada série temporal possui apenas 24 amostras, número este considerado pequeno para o treinamento da RNR. Assim, os dados foram interpolados, utilizando *Splines* [Chambers and Hastie 1991], e cada série passou a ter 48 amostras.

A matriz de entrada da RNR é composta de 9 entradas, uma para cada padrão, com 752 amostras, resultantes das 16 seqüências de amostras, interpoladas e postas de maneira seqüencial (conforme descrito em maiores detalhes na Seção 5.2.2).

Uma vez composta a base de treinamento, a próxima etapa é o treinamento da RNR, que será descrito na seção seguinte.

6.2.4 Treinamento da RNR

Os dados deste experimento são mais complexos quando comparados aos do experimento anterior, desta forma, exigindo uma rede mais complexa para modelá-los. Foram analisadas 4 configurações de RNRs, com 1, 2, 3 e 4 neurônios na camada oculta. Os resultados do treinamento e o número de estados extraídos de cada rede estão apresentados na Tabela 6.2.

TABELA 6.2 – Comparativo de desempenho das RNR treinadas com os dados do CDC-15.

Nro. de neurônios na camada oculta	Número possível de estados	Número de estados extraídos	Erro RMS
1	3	3	0.6422127
2	9	4	0.6291172
3	27	9	0.6141776
4	81	20	0.6090115

O erro RMS do treinamento das RNRs mostrou-se sensível ao aumento no número de neurônios, diminuindo conforme adiciona-se neurônios na camada oculta.

Quanto ao número de regras extraídas, a rede com apenas 1 neurônio determinou o número máximo de 3 estados possíveis. Diferente do experimento anterior, ao aumentarmos o número de neurônios ocultos, aumentou também o número de estados extraídos, logo, uma rede com apenas 1 neurônio não é capaz de *aprender* toda a informação presente nos dados. As outras configurações apresentaram em média 4, 9 e 20 estados.

Frente a estes resultados, pode-se concluir que o uso de uma rede com 2 neurônios, extraindo apenas 4 estados, pode não representar satisfatoriamente o conjunto de dados. A rede com 4 neurônios e 20 estados, não generaliza a informação obtida. Retornando à descrição dos dados, cada série temporal é composta de 24 amostras. Ao se extrair 20 estados, há quase 1 estado para descrever cada dado apresentado à rede, não havendo generalização na descrição e sim, apenas uma representação na forma de grafo.

Desta forma, a rede com 3 neurônios na camada oculta, apresentada na Figura 6.6, se mostra uma boa opção para a análise desta base de dados, com um número de estados plausível para análise, apresentando uma boa generalização dos dados, permitindo uma análise satisfatória do comportamento do CDC-15.

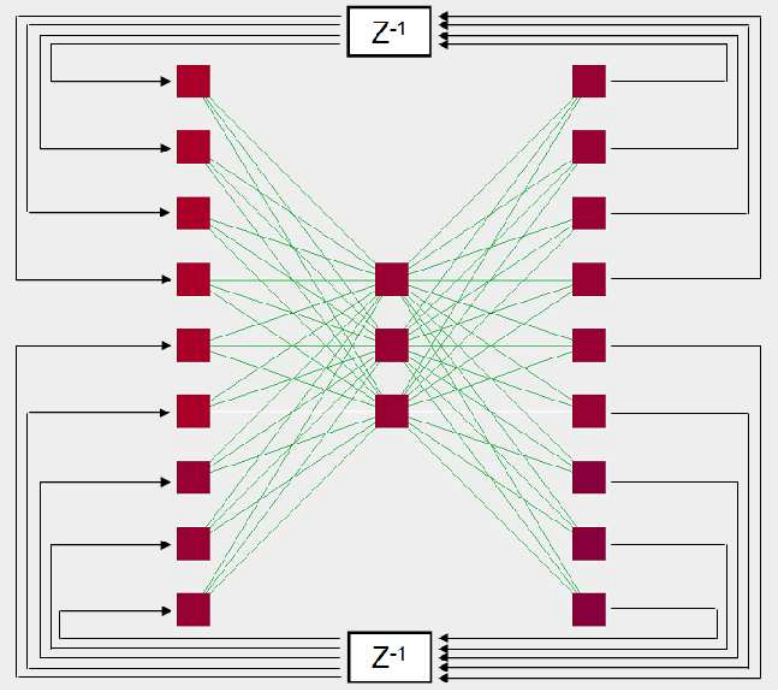


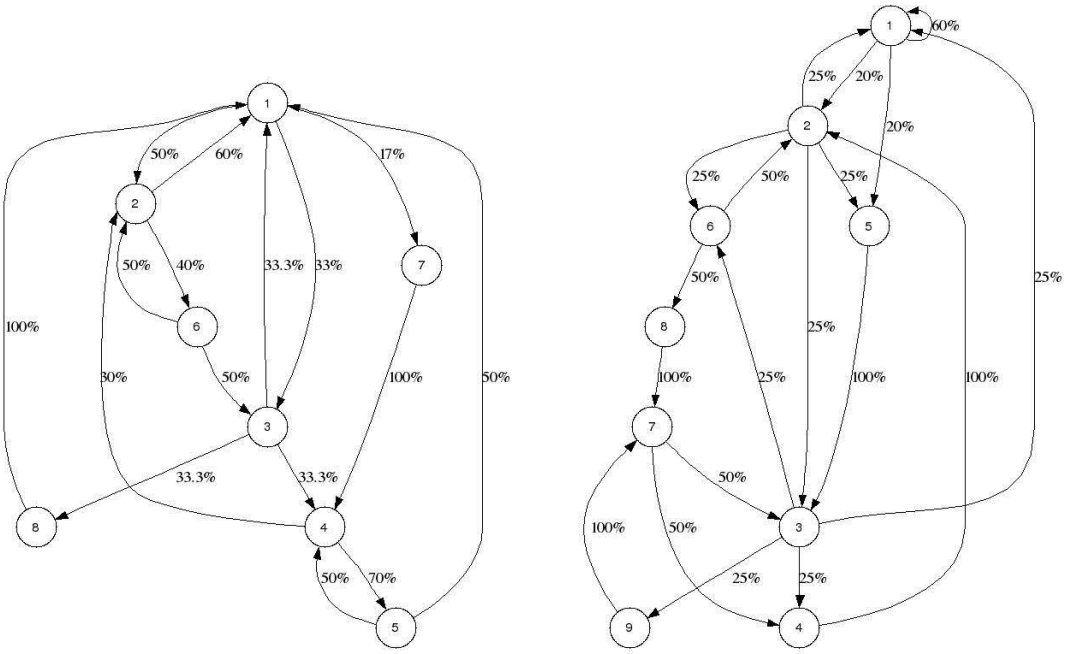
FIGURA 6.6 – *Estrutura da RNR utilizada no experimento com os dados do CDC-15, apresentando uma topologia* $9-3-9$*, baseada na Rede de Jordan.*

A seguir, a próxima Seção descreve a extração da Cadeia de Markov sobre a RNR selecionada.

6.2.5 Extração da Cadeia de Markov

Frente às diferentes configuração de RNRs analisadas, optou-se por realizar a extração sobre a RNR com 3 neurônios na camada oculta. A configuração da Cadeia de Markov extraída varia conforme a RNR apresentada, no entanto as cadeias obtidas mostraram-se bastante semelhantes.

A Figura 6.7 apresenta duas cadeias extraídas de RNRs com 3 neurônios na camada oculta. A cadeia apresentada na Figura 6.7(a) mostra a representação através de 8 estados, e a Figura 6.7(b) com 9 estados. Apesar da diferença entre o número de estados, as cadeias são bem semelhantes, apresentando um comportamento cíclico e a possibilidade de redução.



(a) Cadeia de Markov I - 8 Estados (b) Cadeia de Markov II - 9 Estados

FIGURA 6.7 – *Cadeias de Markov extraídas de RNRs com 3 neurônios na camada oculta, treinadas com dados de expressão gênica do CDC-15.*

A seguir, será apresentada uma breve análise e discussão dos resultados obtidos.

6.2.6 Análise e Discussão dos Resultados

Como mencionado no ínicio da Seção 6.2, o Controle de Divisão Celular é um fenômeno bem estudado pela comunidade científica. Uma das características mais notáveis neste fenômeno é o comportamento cíclico apresentado. A Figura 6.8 ilustra este processo e suas etapas. Maiores informações sobre divisão celular podem ser obtidas em Amabis e Martho [Amabis and Martho 2001].

Para este experimento, a característica fundamental presente nestes dados é o

comportamento cíclico. Sabendo desta característica, o sucesso deste experimento está em obter Cadeias de Markov representando este comportamento.

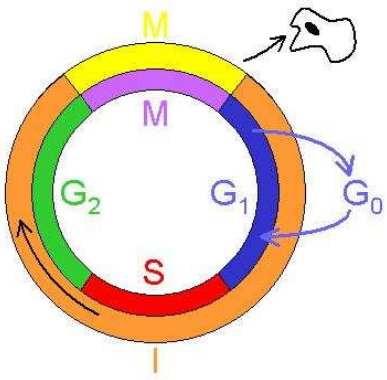


FIGURA 6.8 – *Ciclo de Divisão Celular.*

Para análise foram selecionadas duas cadeias, apresentadas anteriormente na Figura 6.7. Analisando estas Cadeias fica claro o comportamento cíclico apresentado por ambas, pois, a partir de qualquer estado destas cadeias é possível retornar ao mesmo após um certo número t de iterações. Vale salientar que este comportamento cíclico não está explícito nos dados, no entanto, a metodologia pode extrair tal informação.

Outra característica presente nos dados são estados das cadeias que, a princípio, poderiam ser excluídos, como os Estados 7 e 8 da Cadeia I, e os Estados 8 e 9 da Cadeia II. Diz-se que estes estados podem ser excluídos porque apresentam somente transição para um outro estado, apresentando probabilidade de 100%. Para exemplificar, analisaremos o Estado 8 da Cadeia I. Este estado apresenta transição somente para o Estado 1, desta forma, matematicamente este estado poderia ser excluído e haver somente uma transição do Estado 3 para o Estado 1 com uma probabilidade de 33%. Apesar de matematicamente este estado poder ser excluído, vale relembrar o significado deste estado, que é a representação das interações existentes entre os genes em um determinado período de tempo, sendo assim, este estado pode significar uma etapa importante no ciclo celular, sempre precedido de um mesmo fenômeno, neste caso, sempre que ativo o Estado 8, o próximo estado é obrigatoriamente o Estado 1, e excluir o Estado 8 seria perder informações que biologicamente podem ser importantes. No entanto, isto são especulações que precisam ser confirmadas por um especialista no assunto.

Enfim, este experimento mostrou-se importante por apresentar através das Cadeias de Markov extraídas informações coerentes e de acordo com os dados analisados. No entanto, mais estudos são necessários visando uma melhor interpretação dos resultados, mas tal estudo necessitaria do acompanhamento de um especialista no assunto.

A próxima seção apresentará o experimento realizado com a base de dados completa de Stanford.

6.3 Experimento 3 - SMD

Este experimento visa analisar uma base de dados maior e bastante utilizada pelo meio científico, a *Stanford Microarry Database* (SMD)[Ball et al. 2005]. Como dificuldade principal neste experimento, está a complexidade dos dados. São muitas séries amostradas em diferentes fases biológicas. No entanto, existem muitos trabalhos que utilizam estes mesmos dados, o que permite algumas comparações, e também, fonte de informações para futuros estudos, mais aprofundados, sobre os resultados obtidos.

6.3.1 Seleção dos Dados

Além do fato destes dados serem utilizados por diversos pesquisadores em seus trabalhos, outro fator importante que motivou sua utilização é o número de amostras temporais, 80 no total. Normalmente, medições de Microarranjos possuem aproximadamente 20 pontos, assim, frente às demais bases, esta possui um grande número de amostras, facilitando a análise.

A base de dados completa do SMD possui a expressão gênica de mais de 6000 genes. Neste trabalho foi utilizada apenas uma parte desta informação. Foram utilizadas as expressões dos genes selecionados por Eisen *et al* [Eisen et al. 1998], os quais também foram utilizados por Brown *et al* [Brown et al. 2000]. Esta seleção consiste na expressão gênica de 2468 genes, de 6 classes funcionais distintas: Ciclo do Ácido Carboxílico (TCA), Respiração, Ribossomos Citoplasmáticos, Proteassoma, Histonas e Proteínas *Helix-Turn-Helix*. Vale lembrar que cada expressão gênica é uma série temporal composta de 80 amostras.

Quanto ao volume de dados a ser analisado, a base em questão é maior que a analisada no experimento anterior (os dados do CDC apresentavam 799 expressões gênicas), sendo assim, fica evidente a necessidade de uma redução neste volume. Tal processo está descrito em detalhes na próxima seção.

6.3.2 Determinação dos Padrões

Neste experimento foram analisadas 4 configurações para os SOMs, 5×5, 10×10, 15×15 e 20×20. Somente a rede 5×5 não conseguiu separar os dados de maneira satisfatória. As demais redes identificaram 11 padrões principais entre os dados, cujos protótipos mostraram-se muito semelhantes. No entanto, as redes 15×15 e 20×20 dispersaram excessivamente os dados, apresentando uma relação de dados $\times$ padrão significativamente menor do que a rede 10×10.

Quanto ao número de padrões identificados, foi possível comparar estes resultados com os resultados obtidos por Nikkilä *et al* [Nikkila et al. 2002], no qual eles fizeram uso de SOM na análise deste mesmo conjunto de dados buscando encontrar clusters. Segundo seu experimento, foram encontrados 10 grupos entre os dados, fazendo uso de uma rede de dimensões maiores que as utilizadas neste trabalho, número este muito próximo aos padrões determinados.

Os padrões determinados, foram analisados por um especialista para determinar se eles representam algum conjunto conhecido de genes, ou então algum caminho

metabólico. No entanto, não foi possível chegar a nenhuma conclusão sobre o que representa, biologicamente, cada padrão. Tal informação é dificil de ser obtida, uma vez que cada gene pode participar de um ou mais processos biológicos, dificultando a classificação do mesmo ao tentar colocá-lo em uma classe.

Para os padrões determinados neste experimento, cada um apresentou pelo menos 36 dados correspondentes, assim, o próximo passo consiste em compor a base de treinamento para a RNR, o que será descrito a seguir na próxima Seção.

6.3.3 Composição da Base de Treinamento

Seguindo a metodologia, a base de treinamento foi composta pelos dados originais que são representados por cada padrão determinado pelo SOM.

Como foram determinados 11 padrões principais, a RNR terá 11 neurônios na camada de entrada. Assim, cada neurônio corresponderá a um padrão, e receberá como entrada 36 séries de expressões gênicas correspondendo a cada padrão. Como já explicado anteriormente, foi evitada a utilização dos protótipos obtidos com o SOM, uma vez que estes não apresentam tanta informação quanto os dados originais.

Os dados do SMD são séries contendo 80 amostras. Este número de amostras foi considerado satisfatório para a realização dos experimentos, assim, não foi realizado nenhum tipo de pré-processamento a fim de aumentar este número.

Desta forma, a matriz de entrada da RNR é composta de 11 entradas, uma para cada padrão, com 2844 amostras, resultantes das 36 expressões gênicas correspondentes a cada padrão, postas de maneira seqüencial.

6.3.4 Treinamento da RNR

Frente a complexidade e ao volume maior desta base de dados, avaliou-se um número maior de configurações de RNR. Foram analisadas redes com 1, 3, 5, 10, 15 e 20 neurônios na camada oculta. A Tabela 6.3 mostra os resultados obtidos para os experimentos.

TABELA 6.3 – Comparativo de desempenho das RNRs treinadas com os dados do SMD.

Nro. de neurônios na camada oculta	Número possível de estados	Número de estados extraídos	Erro RMS
1	3	2	4.135321
3	27	8	3.956288
5	125	24	3.912200
10	1000	73	3.866480
15	3375	228	3.771372
20	8000	319	3.733685

No entanto, para este caso, determinar qual a rede a ser utilizada na extração da Cadeia de Markov não foi uma tarefa simples. Simplesmente basear a escolha no

erro RMS apresentado pelas redes é uma atitude muito vaga, pois, quanto menor o erro, maior o número de estados.

O número de estados extraídos das redes com 10, ou mais, neurônios ocultos, se torna muito grande. Lembrando que as expressões gênicas deste experimento possuem 80 amostras temporais, no caso da rede com 10 neurônios ocultos, há quase 1 estado para cada amostra temporal, o que torna o resultado muito específico. Quanto a rede com 15 e 20 neurônios a especificidade do resultado do treinamento desce a nível de especificar estados válidos para cada uma das expressões que compõem a base de treinamento, fazendo com que determinados estados sejam válidos para um número muito pequeno de amostras, desta forma, prejudicando a generalização desejada para o treinamento.

Assim, para avaliar as redes treinadas, foram considerados os erros RMS e o número de estados da cadeia extraída e foi feita uma avaliação da capacidade de predição das RNRs. Foram selecionadas entradas aleatórias, que tiveram a predição realizada pelas RNRs avaliadas manualmente. Foi observado que as RNRs com 1 e 3 neurônios não são capazes de gerar uma saída satisfatória. As outras redes tiveram sucesso na predição dos dados.

Após avaliar o erro, o número de estados e a capacidade de predição, a rede com 5 neurônios ocultos foi considerada a melhor para a realização da extração da cadeia. Esta rede pode ser vista na Figura 6.9, sendo a extração da Cadeia de Markov realizada com a mesma, descrita na próxima seção.

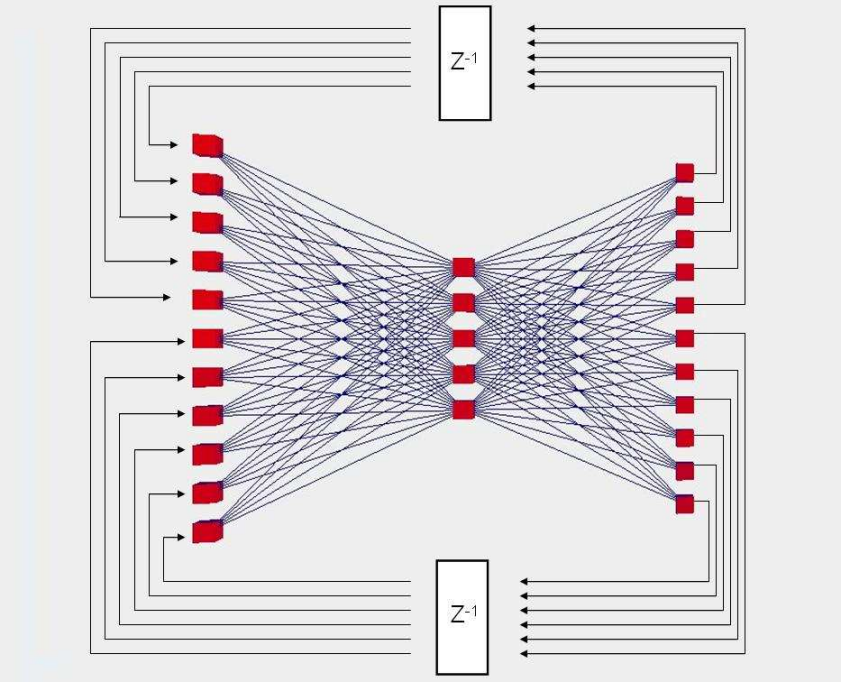


FIGURA 6.9 – *Estrutura da RNR utilizada no experimento com os dados do SMD, apresentando uma topologia* $11 - 5 - 11$*, baseada na Rede de Jordan.*

6.3.5 Extração da Cadeia de Markov

A extração da Cadeia de Markov foi realizada com a RNR com 5 neurônios ocultos. Desta rede foram extraídos 24 estados, dos 125 possíveis. Entretanto, analisando o número de dados representados por cada estado, através de um histograma (ver Figura 6.10), pode-se constatar que apenas os 5 primeiros estados já são suficientes para representar 95% dos dados, ou seja, da informação analisada.

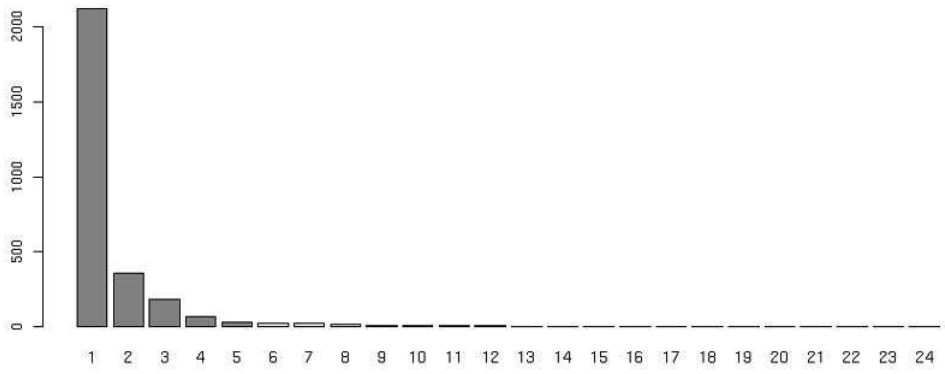


FIGURA 6.10 – *Histograma construído para identificar o número de estados da Cadeia de Markov. O eixo x representa as funções de pertinência, e o eixo y o número de dados representados por elas.*

Em seguida os dados foram clusterizados entre os 5 primeiros estados da Cadeia de Markov. Analisou-se estatisticamente as séries temporais de dados a fim de determinar as trocas de estados existentes, bem como a freqüência com que ocorrem. A Cadeia obtida está apresentada na Figura 6.11.

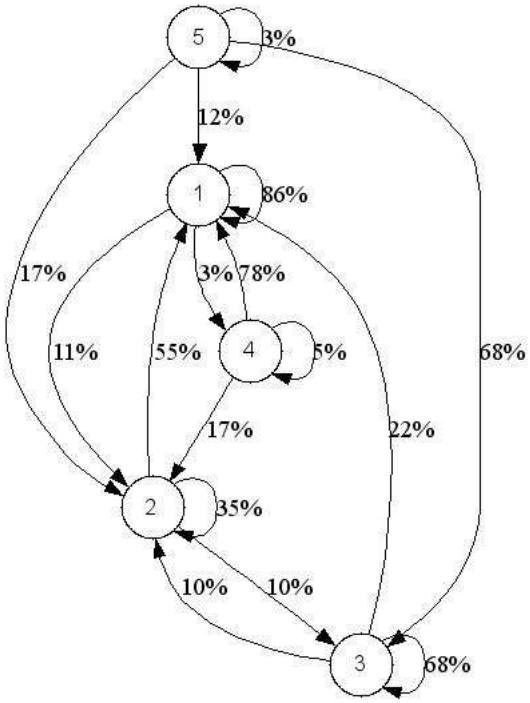


FIGURA 6.11 – *Cadeia de Markov extraída de uma RNR com 5 neurônios na camada oculta, treinada com dados do SMD.*

6.3.6 Análise e Discussão dos Resultados

A SMD é a maior base de dados entre as analisadas nos experimentos apresentados, no entanto chama a atenção o fato de apenas dois estados descreverem praticamente 90% dos dados. Desta forma, considerando os resultados, analisou-se a representatividade dos estados da Cadeia.

Ao analisar a representatividade de cada um dos estados foi observada a rara ocorrência dos estados 3, 4 e 5. Estes estados possuem importância muito pequena quando comparados aos estados 1 e 2, os quais estão bem definidos na série temporal.

A Figura 6.12 mostra as regiões representadas pelos estados 1 e 2 nos dados. Observa-se uma diferença clara entre as regiões representadas e a generalização feita pela RNR sobre os dados. As 60 primeiras amostras, região dos dados com maior ruído, foram praticamente todas atribuídas ao comportamento descrito pelo Estado 1 enquanto que a parte final dos dados (as 20 últimas amostras) foram atribuídas ao segundo estado. Nesta figura os dados representados pelos demais estados não foram mostrados devido a baixa ocorrência e para facilitar o entendimento da explicação.

O Estado 1 é o estado mais representativo sobre os dados, no entanto, analisar as relações existentes na sua região é uma tarefa difícil, frente à pequena amplitude dos sinais e ao grande nível de ruído. No entanto as relações determinadas por este estado foram representadas em forma de grafo e podem ser vistas na Figura 6.13. Neste grafo, cada nodo representa um conjunto de genes, ou seja, uma entrada da RNR, e os arcos, são as relações existentes entre estes genes. Os números junto aos arcos representam estas relações e foram obtidos diretamente dos pesos da RNR.

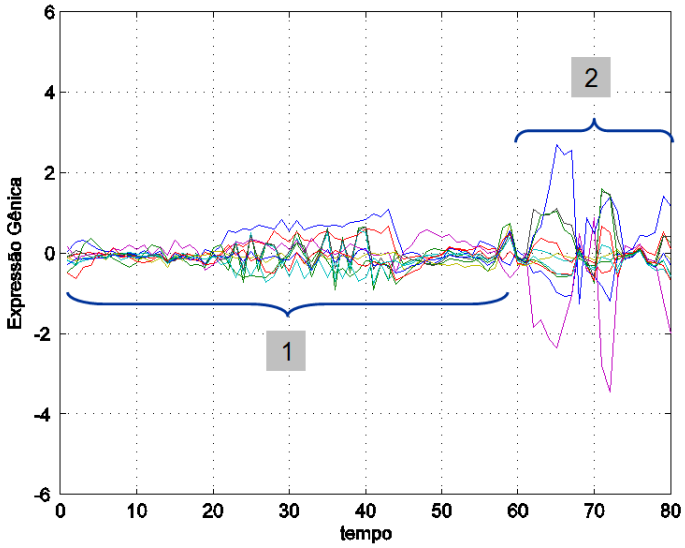


FIGURA 6.12 – *Regiões representadas pelos dois primeiros estados da Cadeia de Markov extraída.*

Seu módulo tem significado relativo entre estes dados, representando qual influência é maior que outra, e o sinal representa se a influência é positiva ou negativa. Dizer que a influência é positiva ou negativa, significa dizer que estes sinais comportam-se de maneiras opostas, enquanto um está crescendo, o outro está decrescendo, e vice-versa.

O Estado 2 representa uma porção menor dos dados, mas nem por isso é considerado um estado menos importante. A região representada por ele apresenta sinais mais definidos, facilitando o estudo. Este estado está representado no grafo apresentado na Figura 6.14.

Analisando o segundo estado, há indícios de haver uma seqüência de eventos nos quais os genes representados pelos grupos C, D, B, I, E e G encontram-se envolvidos. Alguns genes mostram-se como quadjuvantes no processo, como os representados pelos grupos K e F, influenciando alguns grupos, mas aparentemente fora de uma seqüência de eventos. As relações dos genes K e F com os demais, apresentam um módulo elevado, no entanto, ao analisar os genes dos grupos foi visto que os sinais possuem amplitude baixa, dificultando possíveis análises sobre seu comportamento.

Há uma outra região bastante interessante para análise, a região central envolvendo os clusters I, E e G. São observados indícios de regulação entre os 3 clusters, podendo haver regulação do tipo *Feedfoward loop* entre eles, neste caso, a ativação de I e G ocasionaria a repressão dos genes no grupo E.

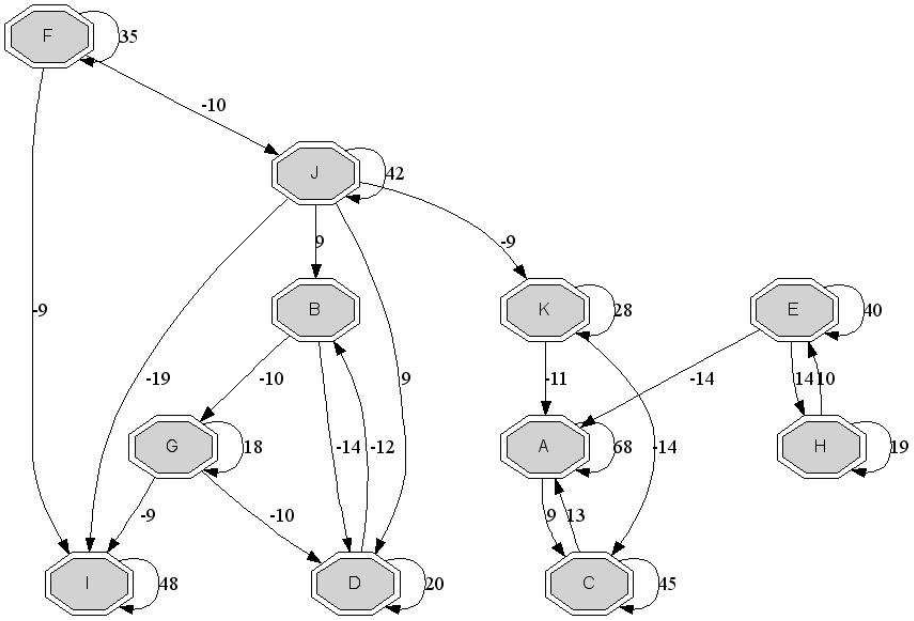


FIGURA 6.13 – *Grafo representando o principal estado (estado 1 na Figura 6.11) da Cadeia de Markov extraída. Os nodos representam os grupos de genes, e os arcos as influências.*

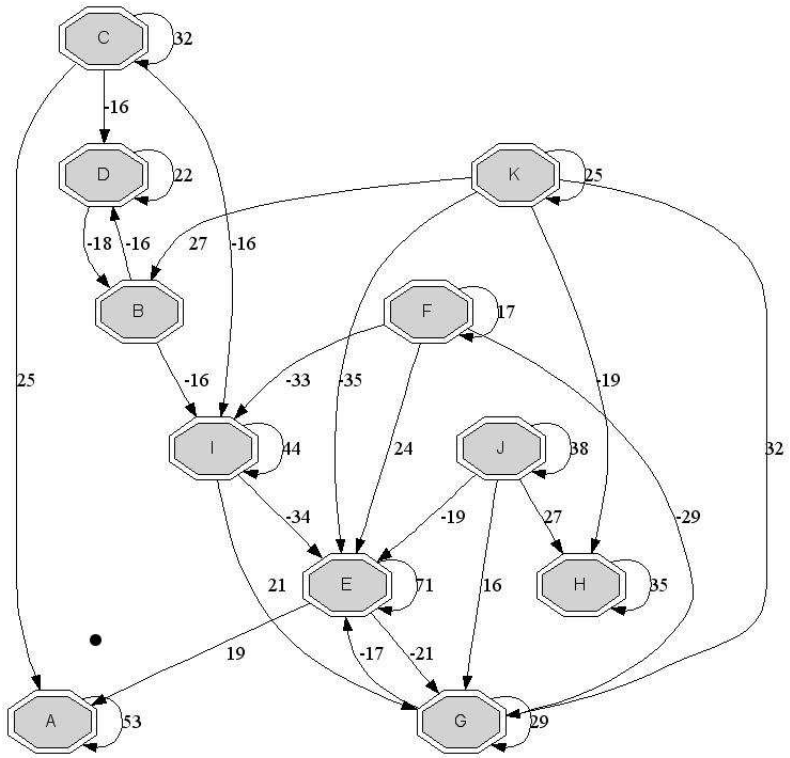


FIGURA 6.14 – *Grafo representando o segundo principal estado (estado 2 na Figura 6.11) da Cadeia de Markov extraída. Os nodos representam os grupos de genes, e os arcos as influências.*

As relações existentes entre os clusters I, G e E (destacada na Figura 6.15) foram analisadas através dos protótipos de seus clusters, visando confirmar as relações apresentadas. Conforme apresentado, há dois tipos de comportamento entre os dois clusters de genes. Os clusters I e G possuem o mesmo comportamento, uma vez que estão ligados com uma aresta de peso positivo. Já as relações de I com E e de E com G são ligações negativas, o que indica que o cluster E comporta-se de maneira oposta a I e G.

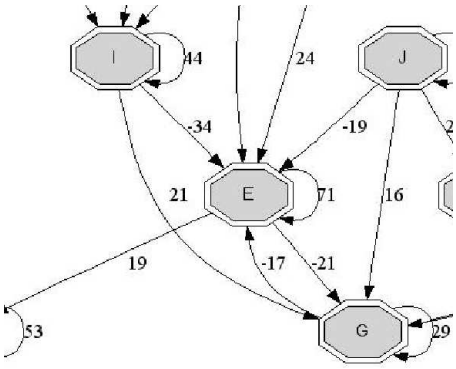


FIGURA 6.15 – *Região do grafo referente as relações apresentadas no Estado 2 (Figura 6.14), a ser estudada em maiores detalhes.*

As hipóteses acima mencionadas necessitam verificação. Isto foi feito analisando os protótipos dos clusters em questão. As vinte últimas amostras temporais destes protótipos, ou seja, os pontos cujo comportamento é representado pelo Estado 2, são apresentadas na Figura 6.16. Ao analisar a figura confirma-se a análise feita anteriormente através do gráfico representando as relações presentes entre os dados no Estado 2. O padrão G (em azul) e o padrão I (em preto), apresentam praticamente o mesmo comportamento, estando, suas curvas, praticamente sobrepostas. Já o padrão E (em vermelho) apresenta um comportamento oposto aos outros dois, desta forma, enquanto o padrão E apresenta um comportamento crescente, G e I decrescem, e vice-versa.

A metodologia proposta mostrou-se eficiente a ponto de determinar em um grande conjunto de dados (no caso, dados de Microarranjos) os padrões mais representativos entre os dados, e através de uma metodologia de extração de conhecimento, determinar de forma qualitativa e quantitativa as relações entre estes dados. No caso de dados de Microarranjos, estes resultados geram indícios de regulação presente entre os genes. No entanto, dizer que estes genes estão realmente participando de um processo de regulação é um tanto arriscado, pois muitos genes podem apresentar um comportamento similar, ou oposto, sem necessariamente estarem participando de um mesmo processo, mas no entanto, estes resultados apresentam, no mínimo, um indício de quais genes podem estar participando do mesmo processo. Frente a um grande conjunto de dados, estes indícios podem ser muito úteis para um especialista que pretende, frente a milhares de expressões gênicas, buscar possíveis relações de regulação.

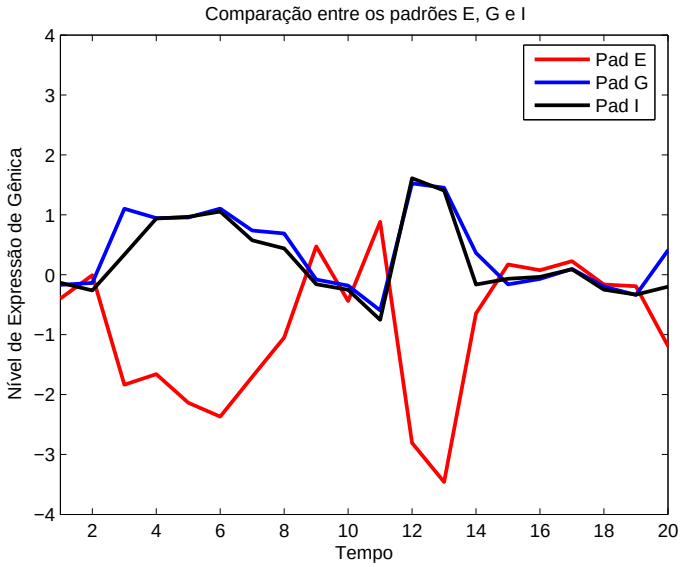


FIGURA 6.16 – *Comparação entre os protótipos dos padrões I, E e G.*

6.4 Experimento 4 - Canais de Íon

Este último experimento analisou os dados fornecidos pelos laboratórios da HP R&D de Palo Alto. São expressões gênicas amostradas dos Canais de Íon dos canais auditivos de ratos [Rivolta et al. 2002]. O interesse destes laboratórios é encontrar genes que participem do processo infeccioso que sofrem os ratos. Determinando os genes que participam de tal processo, é possível buscar maneiras para combater este processo, evitando que a doença se propague.

Os dados em questão são dados de estudos atuais, logo, pouco se sabe sobre eles. Desta forma, toda informação obtida pela análise realizada pode servir de indícios para dirigir as pesquisas realizadas.

6.4.1 Seleção dos Dados

Os dados foram fornecidos em sua forma bruta. Consistem na expressão de 6584 genes com 12 amostras temporais não equidistantes. Assim, analisar tal base de dados é um grande desafio, pois frente a um número pequeno de amostras temporais como este, é difícil inferir as interações temporais que estão ocorrendo.

O primeiro passo foi interpolar os dados, gerando mais amostras temporais para otimizar o treinamento da RNRs. As 12 amostras foram interpoladas em 60 pontos equidistantes. Após a interpolação, os dados foram novamente analisados. Observou-se uma grande diferença no intervalo de valores destas séries. Muitos genes apresentavam seu nível de expressão praticamente nulo durante os experimentos. Outros não apresentavam-se nulos, mas em compensação, não apresentavam grandes variações. Ambos não apresentam informações relevantes e foram excluídos da análise.

Outro problema presente nos dados foi o intervalo de valores apresentado, na ordem de 10^4. Para resolver este problema os dados foram padronizados segundo a equação abaixo:

$$dados\_pad = \frac{dados\_originais - media\,(dados\_originais)}{desv\_pad\,(dados\_originais)} \tag{6.5}$$

ou seja, remove-se a média dos dados, e divide-se pelo desvio padrão. Desta forma o conjunto apresenta média = 0.0 e desvio padrão = 1.0. Maiores informações sobre padronização podem ser obtidas em Law & Kelton [Law and Kelton 2000].

Por fim restaram 771 genes que foram analisados a fim de identificar as relações presentes entre eles.

6.4.2 Determinação dos Padrões

Apesar de, através dos critérios impostos na fase de seleção, ter havido uma significativa redução no número de variáveis a serem analisadas, este número ainda é considerado alto quando o objetivo é o treinamento de redes neurais. Assim, repetiu-se aqui o mesmo procedimento de determinação de padrões citado nos experimentos das Seções 6.2 e 6.3.

Na busca dos padrões presentes nos dados foi utilizado um SOM de topologia 10×10. Submetida ao treinamento, esta rede identificou a presença de 9 principais padrões, cada padrão, representando pelo menos 15 expressões gênicas. Estes padrões são mostrados na Figura 6.17.

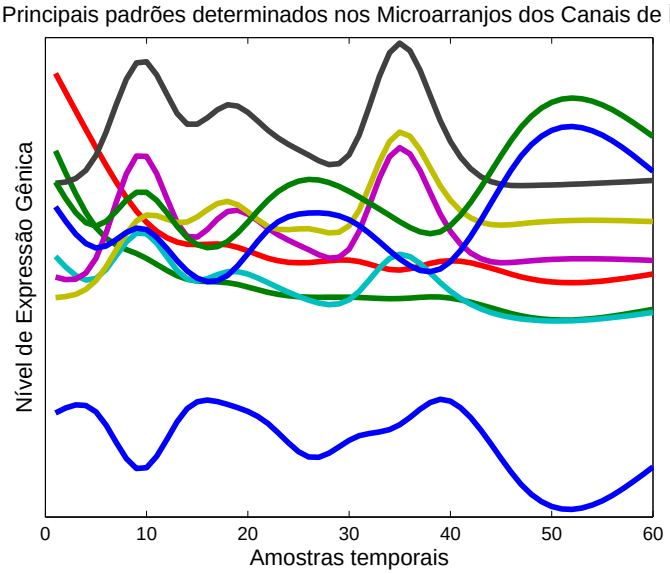


FIGURA 6.17 – *Principais padrões presentes nos Microarranjos de canais de íons determinados por um SOM de* 10×10.

Ao observar estes padrões há a impressão de que alguns deles parecem ser o mesmo padrão, no entanto, com nível de expressão gênica deslocado. Isto merece

um estudo mais detalhado. O comportamento temporal destes genes é o mesmo, no entanto o fato de apresentarem maior ou menor nível de expressão gênica pode representar uma informação biologicamente importante. Como não foi possível a análise destes resultados por um especialista, eles serão considerados como sendo padrões distintos de comportamento.

6.4.3 Composição da Base de Treinamento

A base de treinamento para este experimento foi composta a partir dos 9 padrões determinados pelo SOM. Cada padrão corresponde a um parâmetro de entrada da RNR, que receberá as expressões gênicas correspondentes ao respectivo padrão.

Os dados em questão, após interpolados, possuem 60 amostras temporais. Cada padrão representa pelo menos 15 expressões gênicas. Desta forma, a matriz de entrada da RNR é composta por 9 entradas, cada uma correspondendo a um dos padrões determinados, e 885 amostras, correspondendo as 15 expressões gênicas de cada padrão, que foram postas sequencialmente.

6.4.4 Treinamento da RNR

Para identificar a RNR a ser utilizada, foram analisadas 4 configurações, com 1, 2, 3 e 4 neurônios na camada oculta. A Tabela 6.4 apresenta os resultados obtidos.

TABELA 6.4 – Comparativo de desempenho das RNRs treinadas com os dados dos Canais de Íons.

Nro. de neurônios na camada oculta	Número possível de estados	Número de estados extraídos	Erro RMS
1	3	2	0.003134426
2	9	5	0.003051122
3	27	6	0.002959798
4	81	8	0.002847120

Analisando o erro RMS, observa-se que o aumento no número de neurônios na camada oculta melhora o desempenho da rede. Quanto ao número de estados extraídos, este se mantém pequeno, apresentando pequenas variações de uma configuração para outra. A rede com 4 neurônios mostra uma configuração robusta demais para os dados em questão, pois a mesma representa menos de 10% dos estados que poderia estar representando.

Assim, optou-se por trabalhar com a rede com 3 neurônios na camada oculta, uma vez que a mesma demonstrou um bom aprendizado, e uma boa relação entre estados possíveis de serem extraídos frente aos que realmente foram extraídos. Desta forma, a rede considerada a melhor para este experimento apresenta a mesma topologia da rede do experimento 2 (Seção 6.2). O número de padrões determinados nos dados foi o mesmo, também neste experimento, considera-se que 3 neurônios

ocultos resultam em um bom desempenho da rede. Isto resultou em uma RNR de topologia $9 - 3 - 9$, conforme a rede apresentada na Figura 6.6.

6.4.5 Extração da Cadeia de Markov

A extração da Cadeia de Markov foi realizada com a RNR com 3 neurônios ocultos. Desta rede extraíram-se 6 estados, dos 27 possíveis. Foi feito um histograma dos dados representados por cada estado. Analisando este histograma (ver Figura 6.18) conclui-se que os 3 primeiros estados representam mais de 92% dos dados analisados.

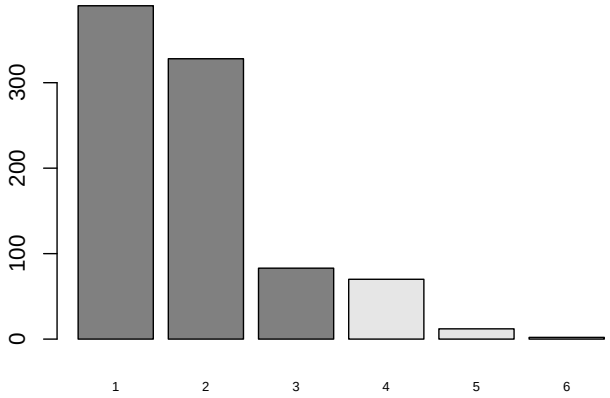


FIGURA 6.18 – *Histograma construído para identificar o número de estados da Cadeia de Markov. O eixo x representa as funções de pertinência, e o eixo y o número de dados representados por elas.*

Frente às conclusões chegadas através da análise do histograma, os dados foram clusterizados entre os 3 primeiros estados. Realizou-se a análise estatística da série temporal de dados para determinar as transições, e obteve-se a Cadeia de Markov apresentada na Figura 6.19.

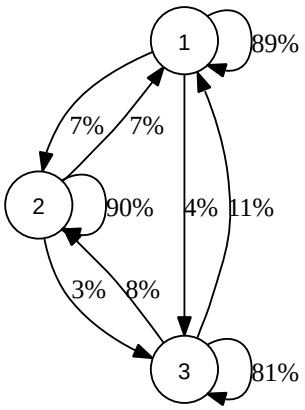


FIGURA 6.19 – Cadeia de Markov extraída dos dados de Canais de Íons.

Quanto as interações descritas nos estados, as Figuras 6.20 e 6.21 apresentam, respectivamente, as interações dos estados 1 e 2 da cadeia de Markov da Figura 6.19.

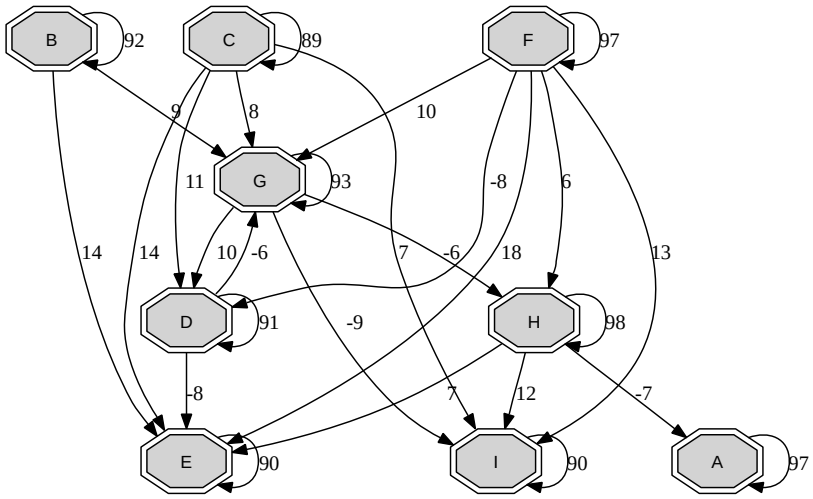


FIGURA 6.20 – *Grafo representando o estado 1 da Cadeia de Markov extraída (Figura 6.19). Os nodos representam os grupos de genes, e os arcos as influências.*

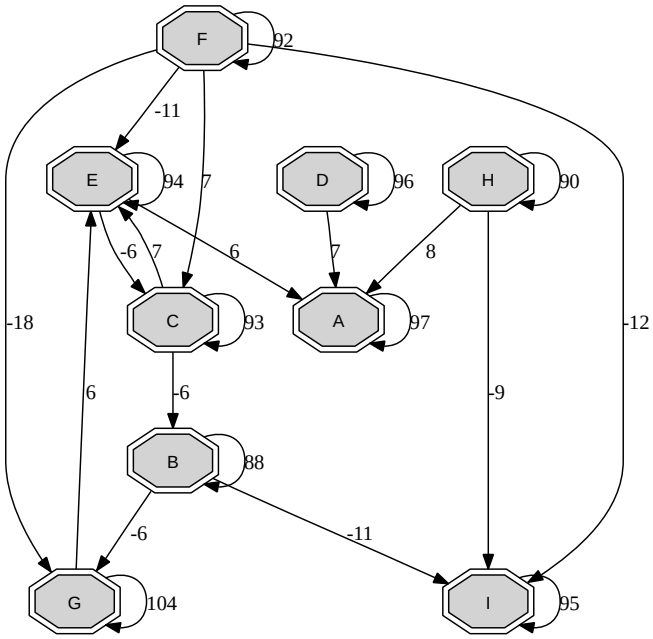


FIGURA 6.21 – *Grafo representando o estado 2 da Cadeia de Markov extraída (Figura 6.19). Os nodos representam os grupos de genes, e os arcos as influências.*

6.4.6 Análise e Discussão dos Resultados

Uma característica, em especial, destes dados, dificultou a análise. A pequena quantidade de amostras temporais. Frente a isto foi necessário interpolar os dados.

Originalmente haviam 12 amostras, e após a interpolação, 60. Tal interpolação pode ter inserido erro aos dados mas, sem ela não seria possível treinar as RNRs e dar seguimento às etapas propostas.

Outra consideração a ser feita é o fato dos dados de Microarranjos não serem precisos, principalmente pelo grande espaço de tempo entre uma amostragem dos dados e outra. Assim, durante este período de tempo, muitas vezes não se pode fazer afirmações sobre o comportamento dos genes. Com isto, os erros inseridos pela interpolação podem não ser tão significativos, no entanto, um estudo mais detalhado é necessário.

Quanto à cadeia extraída, por motivo de tempo não foram analisadas outras configurações. Pela característica dos dados analisados, acredita-se ser possível determinar estados de absorção nas cadeias extraídas. No entanto, com a representação dos dados somente em 3 estados, isto não fica evidente.

As interações entre os genes, descritas pelos estados, permitem observações interessantes. Primeiramente, com base no histograma dos dados, pode-se dizer que existem basicamente dois momentos nos dados, um representado pelo Estado 1 e outro, pelo Estado 2. Alguns grupos de genes apresentam características peculiares. Entre eles cite-se os grupos I e A. Analisando os gráficos das Figuras 6.20 e 6.21, observa-se que estes grupos não influenciam nenhum outro, seja de forma positiva, ou negativa. Eles somente recebem influência dos demais genes.

Os genes do grupo B mostram comportamento distinto nos dois estados analisados. No Estado 1 (Figura 6.20) eles mostram-se como ativadores, influenciando de forma positiva os grupos G e E. Já no Estado 2 (Figura 6.21) sua função muda, e ele se comporta como repressor.

Outro grupo interessente é o F. Ele apresenta, como alguns outros, comportamento, ao mesmo tempo, de indução de alguns grupos e repressão de outros. Analisando o Estado 1 (Figura 6.20), F está induzindo a expressão de G, H e I, e reprimindo D. Pode-se dizer que ele tem um comportamento predominantemente de indução dos demais grupos. Tal comportamento se inverte no Estado 2 (Figura 6.21), onde ele está inibindo a ativação dos grupos E, G e I, e influenciando de forma positiva somente o grupo C.

Enfim, demais análises sobre estes resultados poderiam ser feitas com o auxílio de um especialista. Tem-se como objetivo, repassar estes resultados aos especialistas que forneceram estes dados para a realização dos experimentos, possibilitando que os mesmos possam analisar cuidadosamente estas informações.

6.5 Encerramento do Capítulo

Este Capítulo apresentou os experimentos realizados, explicando as etapas envolvidas, a determinação dos parâmetros da metodologia proposta, e uma breve discussão sobre os resultados obtidos.

Primeiramente, foi analisada uma base de dados gerada especificamente para o experimento. O objetivo era, analisando os dados, mostrar que a rede consegue determinar um conjunto de equações equivalente ao que gerou os dados. Sendo estas

equações um conjunto de equações diferenciais, a rede sendo capaz de determiná-las, comprova-se que a metodologia é capaz de identificar as relações existentes nos dados. E isto foi mostrado através da comparação das equações obtidas pela extração, com as equações que geraram os dados.

O experimento com a base de dados do CDC-15 teve função vital na análise da representação temporal dos resultados. Neste experimento buscava-se uma característica conhecida, os ciclos presentes nos dados, os quais foram determinados pela extração da cadeia.

Este capítulo também apresentou um estudo das relações determinadas pelos estados das cadeias de Markov, no experimento com os dados do SMD. Foi analisado o significado das relações, bem como sua coerência matemática.

Por fim, realizou-se um experimento para analisar uma base de dados com expressões gênicas de ratos. Pouco se sabe sobre esta base, e os dados fornecidos não são os ideais para este tipo de análise. No entanto foram observadas algumas características interessantes sobre o comportamento dos grupos de genes, principalmente quanto às interações presentes.

Pode-se constatar que a metodologia é uma promissora ferramenta para a análise de Microarranjos, gerando informações importantes sobre o comportamento dos genes analisados.

No próximo capítulo, são apresentadas as conclusões deste trabalho e as possibilidades de trabalhos futuros.

Capítulo 7

Conclusão

Nos últimos anos, os avanços nos projetos de seqüenciamento genômico vêm gerando um grande volume de dados. Entre estes dados, estão os níveis de expressão gênicas medidos através dos Microarranjos. Tais dados possuem como característica o grande volume de dados, sendo, por conseqüência disto, de difícil análise.

Este trabalho propõe uma metodologia de extração de conhecimento para determinar relações temporais de causa e conseqüência presentes em dados de Microarranjos, bem como as relações atemporais. O volume de dados presentes nas medições de Microarranjos exigem técnicas computacionais eficientes para determinar estas relações. Relações que são de grande importância, por exemplo, para a indústria farmacêutica, pois podem fornecer informações importantes para desenvolver drogas voltadas ao combate de vírus e bactérias.

Há uma carência de metodologias eficientes para determinar as interações gênicas. Alguns trabalhos propõem a utilização de Redes Bayesianas, no entanto tal abordagem possui grande limitação computacional, além de não conseguir representar o comportamento temporal dos dados.

Existem autores que utilizam análises estatísticas sobre os dados, utilizando métricas como a correlação, para determinar as interações. No entanto, estas análises geram um grande volume de informação, normalmente de difícil interpretação por parte dos especialistas interessados.

Este trabalho propôs uma metodologia de extração de conhecimento, utilizando, como modelo para os dados, as Redes Neurais Recorrentes. A partir das RNRs foi possível extrair as relações existentes entre os dados, e representá-las através de Cadeias de Markov.

A metodologia proposta é nova. Atualmente não existem trabalhos que realizam a análise de Microarranjos buscando suas relações temporais, e, ao mesmo tempo, fornecendo uma descrição linear das interações presentes nos dados analisados. Como forma de validação, optou-se por aplicar a metodologia sobre um conjunto de dados gerados especificamente com esta finalidade, para depois aplicá-la a dados de Microarranjos. A vantagem de gerar os dados a serem analisados é ter o conhecimento das equações que regem o seu comportamento, desta forma, sabendo o conhecimento que deve ser "extraído" durante o experimento.

O experimento com os dados artificiais pode ser considerado simples, mas os

resultados comprovaram que a metodologia é capaz de descrever o comportamento temporal de uma série de dados, bem como a interação existente entre as séries. Os dados analisados consistiam de três séries temporais geradas através de um conjunto de equações. A metodologia foi capaz de separar os dados, antes com um comportamento não-linear, em dois estados principais, de comportamento linear. A definição destes estados, na forma de equação, foi bem próxima às equações que deram origem ao dados. Desta forma, fica clara a capacidade da metodologia em extrair o comportamento temporal dos dados.

Com o sucesso do experimento sobre os dados artificiais, dados de Microarranjos foram analisados. Primeiramente, foi analisada a base de dados com expressões gênicas referentes ao controle de divisão celular, fênomeno amplamente estudado pelo meio científico com uma característica marcante de ser um fenômeno cíclico. Buscou-se verificar se a metodologia seria capaz de identificar estes ciclos, e uma vez analisados os dados, tal comportamento ficou evidente após a extração da Cadeia de Markov. Vários experimentos foram realizados (mas nem todos apresentados neste trabalho), e em todos os resultados estava presente a formação de ciclos na representação do comportamento temporal dos dados, comprovando a eficiência desta metodologia.

A identificação de ciclos nos dados do CDC-15, reforça a confiabilidade da metodologia quanto aos seus resultados. Este ciclo está presente no fenômeno de divisão celular, no entanto, não é uma informação explícita nos dados de Microarranjos. A modelagem destes dados, utilizando RNRs, possibilitou a extração desta informação e sua representação através de Cadeias de Markov.

Parte do sucesso dos resultados obtidos na extração da Cadeia de Markov com os dados do CDC-15 deve-se à etapa de pré-processamento e transformação realizada sobre os dados. Os dados de Microarranjos, atualmente, não favorecem a análise temporal, pois possuem um grande número de variáveis, com um pequeno número de amostras. Espera-se que no futuro, com o avanço da técnica e redução nos custos, consiga-se gerar mais dados de expressão gênica, aumentando o tempo de monitoramento dos organismos e diminuindo o espaçamento entre as amostragens, conseguindo uma descrição mais detalhada de como se comportam os genes, possibilitando se chegar a resultados melhores.

Em outro experimento, realizou-se a análise da SMD, tornando possível a comparação entre os resultados obtidos na etapa de pré-processamento. A técnica de pré-processamento proposta nesta metodologia, apresenta semelhanças com o trabalho de Nikkilä *et al* [Nikkila et al. 2002], difenciando-se nas técnicas de clusterização aplicadas. Nikkilä *et al* aplicam a técnica da *U-matrix* sobre os SOMs para determinar os clusteres presentes nos dados, enquanto no presente trabalho, utilizou-se a freqüência de ativação dos neurônios da rede para determinar os principais padrões. Apesar destas diferenças entre as abordagens, obteve-se resultados similares. Nikkilä determinou 10 clusteres nos dados, enquanto neste trabalho foram identificados 11 principais padrões de expressão gênica.

Não foi possível analisar todas as interações determinadas através da descrição dos estados obtidos, devido às limitações de tempo intrínsecas a uma dissertação de mestrado. Foram analisadas as relações julgadas mais importantes e com uma força

mais significativa frente às demais. Os resultados foram animadores, pois mostraram que a metodologia é capaz de identificar as relações temporais existentes entre os padrões de dados de cada entrada da RNR.

A identificação de tais relações em dados de Microarranjos gera indícios de interações gênicas de grande importância para pesquisadores desta área. Os resultados não permitem afirmar que a rede de interação é, de fato, a rede regulatória que controla o sistema analisado, mas as relações matemáticas de causa e conseqüência encontradas já são um caminho para isto. Estas relações, determinadas entre os conjuntos de genes, são muito similares às existentes entre os genes que participam de um processo regulatório. Tais resultados mostraram-se matematicamente coerentes, no entanto, são apenas indícios de possíveis regulações, e para confirmar biologicamente estes indícios é necessário um estudo minucioso de um especialista.

Este trabalho apresentou uma nova abordagem para determinação de redes regulatórias. A metodologia, como um todo, apresentou resultados satisfatórios. A etapa de pré-processamento proposta mostrou-se eficaz, permitindo o sucesso do modelo escolhido para o aprendizado. As RNRs foram capazes de modelar satisfatoriamente as séries de expressões gênicas, assim, permitindo a extração da informação armazenada, a qual foi representada atravéz de Cadeias de Markov. Tal representação permitiu uma análise temporal dos dados, importante quando busca-se a melhor compreensão dos organismos estudados. Quanto aos estados das Cadeias de Markov, eles representaram com sucesso as relações temporais presentes nos genes analisados, descrevendo na forma de relações lineares, em seus estados, as interações entre eles.

7.1 Trabalhos Futuros

Este trabalho deixa aberta algumas possibilidades de trabalhos futuros:

- **Simulação.** Cadeias de Markov são uma poderosa ferramenta de simulação, no entanto, neste trabalho elas somente foram utilizadas como um formalismo para representação do conhecimento. Um estudo sobre a capacidade de simulação das cadeias extraídas poderia:
  - Validar as Cadeias de Markov;
  - Proporcionar uma melhor visualização dos dados de expressão gênica;
  - Utilizar a capacidade de simulação como métrica de avalição dos resultados obtidos.
- **Pré-processamento.** Apesar de acreditar que o método de pré-processamento adotado produz bons resultados, outras metodologias poderiam ser analisadas.
- **Parâmetros do sistema**. A metodologia é sensível às variações nos parâmetros, assim, um estudo detalhado que auxilie na escolha dos parâmetros seria interessante.

- **Automatização.** A metodologia é composta de diversas etapas. Estas etapas poderiam ser integradas na forma de um *Framework*, onde algumas outras funcionalidades poderiam ser adicionadas, entre elas: a escolha da técnica de pré-processamento a ser utilizada, e a escolha dos parâmetros, citada anteriormente.

- **Análise por um especialista.** Espera-se, com este trabalho, ter fornecido uma ferramenta para a análise de dados de Microarranjos. No entanto, existe a necessidade da análise dos resultados gerados pela aplicação desta metodologia por parte de um especialista, que poderia indicar possíveis problemas, bem como, os pontos fortes do trabalho.

Bibliografia

[Alberts et al. 2004] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2004). *Biologia Molecular da Célula*. Artmed, 4 edition.

[Amabis and Martho 2001] Amabis, J. and Martho, G. (2001). *Conceitos de Biologia*. Moderna.

[Ando and Iba 2004] Ando, S. and Iba, H. (2004). Classification of Gene Expression Profile Using Combinatory Method of Evolutionary Computation and Machine Learning. *Genetic Programming and Evolvable Machines*, 5(2):145–156.

[Andrews and Geva 2002] Andrews, R. and Geva, S. (2002). Rule extraction from local cluster neural nets. *Neurocomputing*, 47(1-4):1–20.

[Ball et al. 2005] Ball, C., Awad, I., Demeter, J., Gollub, J., Hebert, J., Hernandez-Boussard, T., Jin, H., Matese, J., Nitzberg, M., Wymore, F., et al. (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*, 33(1).

[Bar-Joseph et al. 2002] Bar-Joseph, Z., Gifford, D., Jaakkola, T., and Simon, I. (2002). A new approach to analyzing gene expression time series data. *Proceedings of the sixth annual international conference on Computational biology*, pages 39–48.

[Barabási and Oltvai 2004] Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature*, 5:101–113.

[Ben-Dor et al. 1999] Ben-Dor, A., Shamir, R., Yakhini, Z., et al. (1999). Clustering Gene Expression Patterns. *Journal of Computational Biology*, 6(3-4):281–297.

[Bjornsson and Venegas 1997] Bjornsson, H. and Venegas, S. (1997). A manual for EOF and SVD analyses of climatic data. Technical report, Technical Report from Department of Atmospheric and Oceanic Sciences and Center for Climate and Global Change Research, McGill University, Canada.

[Braga et al. 2000] Braga, A. P., Ludermir, T. B., and Carvalho, A. F. (2000). *Redes Neurais Artificiais. Teoria e aplicações*. LTC.

[Brown et al. 2000] Brown, M. P. S., Brundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97 - Part 1:262–267.

[Butte 2002] Butte, A. (2002). The use and analysis of microarray data. *Nature Reviews Drug Discovery*, 1(12):951–960.

[Butte and Kohane 2000] Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 5:418–29.

[Butte et al. 2000] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *PNAS*, 97(22):12182–12186.

[Causton et al. 2003] Causton, H. C., Quackenbush, J., and Brazma, A. (2003). *Microarray. Gene Expression Data Analysis, A Beginner's Guide*. Blackwell Publishing.

[Cechin 1997] Cechin, A. L. (1997). *The Extraction of Fuzzy Rules from Neural Networks*. PhD thesis, Tubingen, Univ.

[Chambers and Hastie 1991] Chambers, J. and Hastie, T. (1991). *Statistical Models in S*. CRC Press, Inc. Boca Raton, FL, USA.

[Chen et al. 1999] Chen, T., He, H. L., and Church, G. M. (1999). Modeling gene expression with diferential equations. *Pac Symp Biocomput*, pages 29–40.

[Chuang et al. 2004] Chuang, H., Liu, H., Brown, S., McMunn-Coffran, C., Kao, C., and Hsu, D. (2004). Identifying significant genes from microarray data. In *IV IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04)*. IEEE Computer Society.

[Chuang et al. 2003] Chuang, H., Tsai, H., Tsai, Y., and Kao, C. (2003). Ranking genes for discriminability on microarray data. *Journal of Information Science and Engineering*, 19(6):953–966.

[D'haeseleer 2005] D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, 23:1499–1501.

[Duggan et al. 1999] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling usind cdna microarrays. *Nature*.

[Eisen et al. 1998] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868.

[Elman 1990] Elman, J. (1990). Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

[Fayyad et al. 1996a] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996a). *Advances in knowledge discovery and data mining*, chapter From data mining to knowledge discovery: an overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Fayyad et al. 1996b] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996b). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34.

[Frasconi et al. 1996] Frasconi, P., Gori, M., Maggini, M., and Soda, G. (1996). Representation of finite state automata in Recurrent Radial Basis Function networks. *Machine Learning*, 23(1):5–32.

[Freeman and Skapura 1992] Freeman, J. and Skapura, D. (1992). *Neural Networks. Algotirhms, Applications and Programming Techniques.* Addison-Wesley.

[Friedman 2004] Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science's STKE*, 303(5659):799–805.

[Giles et al. 1991] Giles, C., Chen, D., Miller, C., Chen, H., Sun, G., and Lee, Y. (1991). Second-order recurrent neural networks for grammatical inference. *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on*, 2.

[Giles et al. 1992] Giles, C., Miller, C., Chen, D., Chen, H., Sun, G., and Lee, Y. (1992). Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation*, 4(3):393–405.

[Goebel and Gruenwald 1999] Goebel, M. and Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl.*, 1(1):20–33.

[Golub et al. 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 282:531–537.

[Haykin 2001] Haykin, S. (2001). *Redes Neurais. Princípios e prática.* Bookman.

[Hillier and Lieberman 1995] Hillier, F. and Lieberman, G. (1995). *Introduction to operations research.* Holden-Day, Inc. San Francisco, CA, USA, sixth edition.

[Hilsenbeck et al. 1999] Hilsenbeck, S., Friedrichs, W., Schiff, R., O'Connell, P., Hansen, R., Osborne, C., and Fuqua, S. (1999). Statistical analysis of array

expression data as applied to the problem of tamoxifen resistance. *J Natl Cancer Inst*, 91(5):453–9.

[Holter et al. 2000] Holter, N., Mitra, M., Maritan, A., Cieplak, M., Banavar, J., and Fedoroff, N. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS*, 97 (15):8409–8414.

[Imperatore and Bentjerodt 1970] Imperatore, E. and Bentjerodt, R. (1970). *Introduccion a Cadenas de Markov y Programacion Dinamica.* Fac. Ciencias Económicas, Santiago De Chile: Universidad De Chile.

[Isaacson and Madsen 1976] Isaacson, D. and Madsen, R. (1976). *Markov Chain Theory and Applications.* Wiley: New York.

[Jacobsson 2005] Jacobsson, H. (2005). Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review. *Neural Computation*, 17(6):1223–1263.

[Jacobsson and Ziemke 2003] Jacobsson, H. and Ziemke, T. (2003). Reducing complexity of rule extraction from prediction RNNs through domain interaction. Technical report, Tech. Rep. No. HS-IDA-TR-03-007). Skovde: Department of Computer Science, University of Skovde, Sweden.

[Karlin and Taylor 1975] Karlin, S. and Taylor, H. (1975). *A first course in stochastic processes.* Academic Press New York, 2 edition.

[Kim et al. 2003] Kim, S. Y., Imoto, S., and Miyano, S. (2003). Inferring gene networks from time series microarray data using dynamic bayesian networks. *Brief Bioinform*, 4:228–235.

[Kohane et al. 2003] Kohane, I., Kho, A., and Butte, A. (2003). *Microarrays for an integrative genomics.* MIT Press.

[Kohonen 1997] Kohonen, T. (1997). *Self-organizing maps.* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Kohonen et al. 1996] Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1996). SOM PAK: The Self-Organizing Map Program Package. *Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science.*

[Kolen and Kremer 2001] Kolen, J. and Kremer, S. (2001). *A Field Guide to Dynamical Recurrent Networks.* Wiley-IEEE Press.

[Kuo et al. 2004] Kuo, W., Kim, E., Trimarchi, J., Jenssen, T., Vinterbo, S., and Ohno-Machado, L. (2004). A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics*, 37(4):293–303.

[Law and Kelton 2000] Law, A. and Kelton, W. (2000). *Simulation Modeling and Analysis*. McGraw-Hill Higher Education, 3 edition.

[Lehninger et al. 1995] Lehninger, A. L., Nelson, D. L., and Cox, M. M. (1995). *Princípios de Bioquímica*. Sarvier, São Paulo, second edition.

[Liang and Kelemen 2002] Liang, Y. and Kelemen, A. (2002). Mining heterogeneous gene expression data with time lagged recurrent neural networks. *SIGKDD'02*.

[Liu et al. 2004] Liu, Y., Liu, H., Zhang, B., and Wu, G. (2004). Extraction of if-then rules from trained neural network and its application to earthquake prediction. *Cognitive Informatics, 2004. Proceedings of the Third IEEE International Conference on*, pages 109–115.

[Lukashin and Fuchs 2001] Lukashin, A. V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17:405–414.

[McCulloch and Pitts 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

[Mendel and McLaren 1970] Mendel, J. and McLaren, R. (1970). Adaptive, Learning and Pattern Recognition Systems: Theory and Applications, chapter Reinforcement learning control and pattern recognition systems.

[Murphy and Mian 1999] Murphy, K. and Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks. *University of California, Berkeley*.

[Nikkila et al. 2002] Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., and Wong, G. (2002). Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15(8):953–966.

[Noble 2004] Noble, W. S. (2004). *Kernel methods in computational biology*, chapter Support Vector Machine applications in computational biology. MIT Press.

[Noman and Iba 2005] Noman, N. and Iba, H. (2005). Inference of gene regulatory networks using s-system and differential evolution. *Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 439–446.

[Nurse et al. 1998] Nurse, P., Masui, Y., and Hartwell, L. (1998). Understanding the Cell Cycle. *Nature Medicine*, 4(10):1103–1106.

[Oja et al. 2002] Oja, M., Nikkila, J., Toronen, P., Wong, G., Castren, E., and Kaski, S. (2002). Exploratory clustering of gene expression profiles of mutated yeast strains. *Computational and Statistical Approaches to Genomics*.

[Omlin and Giles 1996] Omlin, C. and Giles, C. (1996). Extraction of rules from discrete-time recurrent neural networks. *Neural Networks*, 9(1):41–52.

[Ong et al. 2002] Ong, I., Glasner, J., and Page, D. (2002). Modelling regulatory pathways in E. coli from time series expression profiles. *Bioinformatics*, 18(1):241–248.

[Paul and Iba 2005] Paul, T. and Iba, H. (2005). Extraction of informative genes from microarray data. *Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 453–460.

[Pechmann 2004] Pechmann, D. R. (2004). Extração de conhecimento a partir de redes neurais recorrentes. Master's thesis, UNISINOS.

[Pechmann and Cechin 2004] Pechmann, D. R. and Cechin, A. L. (2004). Representação do comportamento temporal de redes neurais recorrentes em cadeias de markov. In *VIII Brazilian Symposium on Neural Neural Networks (SBRN2004)*, volume 1, pages 1–10.

[Pechmann and Cechin 2005] Pechmann, D. R. and Cechin, A. L. (2005). Comparison of deterministic and fuzzy finite automata extraction methods from jordan networks. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, pages 437–444.

[Perrin et al. 2003] Perrin, B. E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and D'Alche-Buc, F. (2003). Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19:II138–II148.

[Pomeroy et al. 2002] Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442.

[Qian et al. 2003] Qian, J., Lin, J., Luscombe, N., Yu, H., and Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19(15):1917–1926.

[R Development Core Team 2005] R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Ramaswamy et al. 2001] Ramaswamy, R., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al. (2001). Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures. *Proceeding of the National Academy of Sciences of the United States of America*, 98(26):15149–15154.

[Rauber 1998] Rauber, T. W. (1998). Redes neurais artificiais. In *ERI'98 - Encontro Regional de Informática*, pages 201–228, Nova Friburgo-RJ e Vitória-ES.

[Raychaudhuri et al. 2000] Raychaudhuri, S., Stuart, J., and Altman, R. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 5:452–463.

[Rivolta et al. 2002] Rivolta, M., Halsall, A., Johnson, C., Tones, M., and Holley, M. (2002). Transcript Profiling of Functionally Related Groups of Genes During Conditional Differentiation of a Mammalian Cochlear Hair Cell Line. *Genome Research*, 12(7):1091.

[Rumelhart et al. 1986] Rumelhart, D., Hintont, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

[Sakamoto and Iba 2001] Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatory network by using genetic programming. *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, 1:720–726.

[Servan-Schreiber et al. 1989] Servan-Schreiber, D., Cleeremans, A., and McClelland, J. (1989). Learning sequential structure in simple recurrent networks. *Advances in neural information processing systems 1 table of contents*, pages 643–652.

[Servan-Schreiber et al. 1991] Servan-Schreiber, D., Cleeremans, A., and Mcclelland, J. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7(2):161–193.

[Shen-Orr et al. 2002] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation. *Nature Genetics*, 31:64–68.

[Siegelmann and Sontag 1992] Siegelmann, H. and Sontag, E. (1992). On the computational power of neural nets. *Journal of Computer and System Sciences*, 50:132–150.

[Slonim 2002] Slonim, D. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics Supplement*, 32.

[Spellman et al. 1998] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

[Styczynski and Stephanopoulos 2005] Styczynski, M. and Stephanopoulos, G. (2005). Overview of computational methods for the inference of gene regulatory networks. *Computers & Chemical Engineering*, 29(3):519–534.

[Tamayo 1999] Tamayo, P. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS*, 96:2907–2912.

[Tino and Vojtek 1998] Tino, P. and Vojtek, V. (1998). Extracting Stochastic Machines from Recurrent Neural Networks Trained on Complex Symbolic Sequences. *Neural Network World*, 8(5):517–530.

[Toronen et al. 1999] Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett*, 451(2):142–6.

[Ultsch 1993] Ultsch, A. (1993). Self-organizing neural networks for visualization and classification. *Information and Classification*, page 307313.

[Vahed and Omlin 2004] Vahed, A. and Omlin, C. (2004). A Machine Learning Method for Extracting Symbolic Knowledge from Recurrent Neural Networks.

[Voet et al. 2000] Voet, D., Voet, J. G., and Pratt, C. W. (2000). *Fundamentos de Bioquímica*. Artmed, Porto Alegre.

[Wahde and Szallasi 2006] Wahde, M. and Szallasi, Z. (2006). A survey of methods for classification of gene expression data using evolutionary algorithms. *Expert Review of Molecular Diagnostics*, 6(1):101–110.

[Watkins 1991] Watkins, D. (1991). *Fundamentals of matrix computations*. John Wiley & Sons, Inc. New York, NY, USA.

[Watrous and Kuhn 1992] Watrous, R. and Kuhn, G. (1992). Induction of finite-state languages using second-order recurrent networks. *Neural Computation*, 4(3):406–414.

[Yeger-Lotem et al. 2004] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcriptional-regulation and protein-protein interaction. *PNAS*, 101:5934–5939.

[Yeung and Ruzzo 2001] Yeung, K. and Ruzzo, W. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.

[Zeng et al. 1993] Zeng, Z., Goodman, R., and Smyth, P. (1993). Learning finite machines with self-clustering recurrent networks. *Neural Computation*, 5(6):976–990.

[Zhao and Zaki 2005] Zhao, L. and Zaki, M. (2005). TRICLUSTER: an effective algorithm for mining coherent clusters in 3D microarray data. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 694–705.

[Zhou et al. 2004] Zhou, X., Wang, X., Pal, R., Ivanon, I., Bittner, M., and Dougherty, E. R. (2004). A bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. *Bioinformatics*, 20(17):2918–2927.

[Zhou 2004] Zhou, Z. (2004). Rule extraction: using neural networks or for neural networks? *Journal of Computer Science and Technology*, 19(2):249–253.

[Zou and Conzen 2005] Zou, M. and Conzen, S. D. (2005). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79.