UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO INDUSTRIAL

APLICAÇÕES DE COMPUTAÇÃO BIOINSPIRADA EM BIOINFORMÁTICA: INVESTIGANDO O PAPEL DOS GENES E SUAS INTERAÇÕES

George Barreto Pereira Bezerra

Orientador: Prof. Dr. Fernando José Von Zuben
DCA/FEEC/Unicamp

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Área de Concentração: Engenharia de Computação

Campinas – São Paulo – Brasil
Julho de 2006

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

Bezerra, George Barreto Pereira
B469a Aplicações de computação bioinspirada em bioinformática: investigando o papel dos genes e suas interações / George Barreto Pereira Bezerra. --Campinas, SP: [s.n.], 2006.

Orientador: Fernando José Von Zuben
Dissertação (Mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Redes gênicas reguladoras. 2. Expressão gênica. 3. Bioinformática. 4. Osciladores biológicos. 5. Identificação de sistemas I. Von Zuben, Fernando José. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Applications of bioinspired computing in bioinformatics: analyzing the role of genes and their interactions.
Palavras-chave em Inglês: Genetic regulatory networks, Gene expression, Bioinformatics, Biological oscillators, Systems identification.
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Fernando José Von Zuben, Gustavo Maia Souza, Márcio Luiz de Andrade Netto e Rafael Santos Mendes.
Data da defesa: 31/07/2006

Banca examinadora

Fernando José Von Zuben

Fernando José Von Zuben (DCA/FEEC/Unicamp) – Presidente

Gustavo Maia Souza (UNOESTE/SP)

Márcio Luiz de Andrade Netto (DCA/FEEC/Unicamp)

Rafael Santos Mendes (DCA/FEEC/Unicamp)

Dedico esse trabalho aos meus pais, Agildo e Gisele, e aos meus irmãos, Marcelo e Eduardo.

Agradeço

Ao grande apoio e amizade de todos os moradores da república “Lá Ele”: Rodrigo, Maurélio, Ricardo, Lourenço, Gian, Elicarlos, Sérgio, Fernando, Júlio, Guilherme, Tiago (b4), Thiago (Manga) e Giuliano.

A todos os meus amigos do LBiC, sempre companheiros no trabalho e na brincadeira: Helder, Tiago, Wilfredo, Eurípedes, Pablo, Hamilton, Marcelo, Renato, Renan, Mariana e Patrícia.

Ao meu orientador Fernando Von Zuben e ao Prof. Leandro Nunes de Castro, verdadeiros mestres para mim.

À comunidade baiana na Unicamp, pelo suporte cultural, muito importante durante minha vida em Campinas.

A Fernanda e a Lara, que sempre me deram muito apoio e com quem aprendi muito.

Resumo

Esta dissertação trata das redes gênicas, o mecanismo de controle da ativação dos genes nas células, sob três perspectivas computacionais diferentes. Inicialmente, sob uma ótica de engenharia, é elaborada uma ferramenta de inferência de redes gênicas, capaz de reconstruir a estrutura estática dessas redes a partir de um conjunto de dados experimentais. O método proposto para essa tarefa de identificação de sistemas é especialmente projetado para conjunto de dados reduzidos, um cenário bastante comum quando se trata de dados de expressão gênica. Numa segunda etapa, é proposto um modelo computacional das redes gênicas, em que as reações bioquímicas que ocorrem na célula são vistas como equações não-lineares arranjadas numa estrutura conexionista. Desta vez, ao invés de inferir redes existentes, esse modelo é utilizado em conjunto com uma abordagem evolutiva para sintetizar redes gênicas artificiais capazes de realizar tarefas dinâmicas – em específico, para solucionar um problema clássico de robótica evolutiva. Embora o modelo seja empregado como técnica de resolução de problemas, o objetivo agora é mais no sentido científico, isto é, as redes gênicas artificiais evoluídas são analisadas como modelos que podem ajudar a compreender propriedades observadas nos sistemas naturais. Finalmente, a terceira etapa consiste numa abordagem conceitual. O propósito principal é tentar compor um novo cenário para o estudo das redes gênicas, reunindo conceitos e dados empíricos de outras áreas da ciência moderna, como a neurociência e a sinergética, e investigando as implicações de uma nova ótica para o processamento de informação celular. O objetivo aqui é voltado para a compreensão dos mecanismos de processamento de informação em organismos vivos.

Abstract

This dissertation deals with genetic networks, the mechanism of control of gene activity in cells, under three different computational perspectives. Initially, as an engineering approach, a computational tool for inference of genetic networks is proposed, which is able to recover the static structure of these networks from experimental datasets. This systems identification method is especially designed for small datasets, a common scenario when coping with gene expression data. In the second step, a computational model for genetic networks is proposed, in which biochemical reactions that occur inside the cell are treated as nonlinear equations in a connectionist structure. Rather than inferring networks from data, this model is used together with an evolutionary algorithm to synthesize artificial genetic networks that are able to solve dynamic tasks – and in particilar, to solve a classic problem in evolutionary robotics. Although the model is used as a problem-solving technique, the objective here is primarily scientific, i.e., the evolved artificial genetic networks are viewed as an opportunity to study properties observed in natural systems. Finally, the third step comprises a conceptual approach, in which ideas from other fields of modern science, like neuroscience and synergetics, are put together to compose a new scenario to the study of the information processing in genetic networks.

Índice

Este trabalho foi desenvolvido com suporte financeiro do
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Capítulo 1

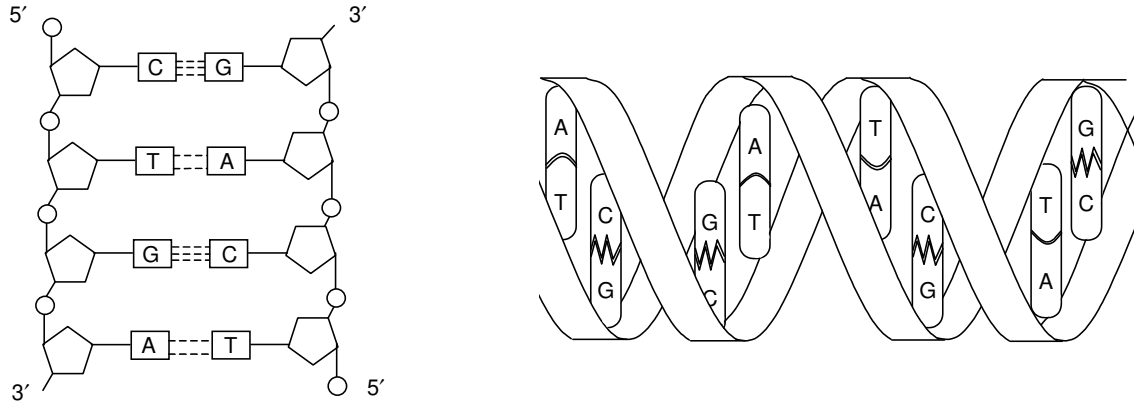
Introdução às Redes Gênicas

Resumo – Este capítulo traz uma introdução a conceitos básicos relativos às redes gênicas. A exposição desses conceitos é breve e suprime detalhes muito específicos de forma a enfatizar os aspectos mais relevantes para a compreensão dos capítulos ulteriores. A Seção 1.1 apresenta uma descrição das unidades básicas do sistema, como DNA, genes e proteínas. A Seção 1.2 explica o processo de expressão gênica e como a expressão pode ser medida e convertida em valores numéricos. A Seção 1.3 introduz o conceito de regulação gênica e como são constituídas as redes reguladoras. As técnicas de modelagem de redes gênicas mais utilizadas na literatura são discutidas na Seção 1.4, e a Seção 1.5 apresenta alguns dados relativos à estrutura dessas redes.

1.1. Conceitos Básicos

A. DNA e RNA

O DNA (ácido desoxirribonucléico) consiste em duas longas fitas, cada uma composta de unidades chamadas fosfatos, moléculas de açúcar e nucleotídeos, ligados em série, formando estruturas denominadas bases nucleotídicas. Existem quatro tipos de nucleotídeos possíveis no DNA: adenina (A), guanina (G), citosina (C) e timina (T). Para facilitar a visualização, é conveniente representar as moléculas de DNA simplesmente por uma seqüência de símbolos correspondentes às bases nucleotídicas da fita {A,G,C,T}. As duas fitas de DNA se encontram ligadas através de pontes de hidrogênio entre suas bases nucleotídicas, segundo regras de paridade, nas quais adenina se liga apenas com a timina (A – T), e a guanina apenas com a citosina (G – C), formando uma estrutura em dupla hélice. Dessa maneira, as fitas de DNA são exatamente complementares entre si. A Figura 1.1 mostra um esquema da molécula de DNA.



(a) (b)

Figura 1.1 Esquema da molécula de DNA. (a) Destaque para as bases nucleotídicas e suas pontes de hidrogênio. (b) Estrutura em dupla hélice. (Fonte: DE CASTRO, 2006)

A molécula de RNA (ácido ribonucléico) é composta por uma fita única. Ela é produzida de forma a complementar uma das fitas do DNA, sendo que seus nucleotídeos são adenina (A), guanina (G), citosina (C) e uracila (U) (este último substitui a timina).

B. Genes

Os genes (ALBERTS *et al*., 1989) são as unidades informacionais básicas da hereditariedade. Eles são seqüências específicas de bases nucleotídicas, as quais carregam as informações necessárias para a construção de proteínas, responsáveis pelos componentes estruturais das células, tecidos e enzimas. Cada molécula de DNA contém vários genes. O conjunto de todos os genes do DNA de um organismo é chamado genoma.

Em um gene existem regiões (seqüências) que dão origem a produtos que exercem propriedades funcionais (exons) e regiões que simplesmente não codificam nenhum produto (íntrons). Acredita-se, porém, que os íntrons possuem um papel muito importante no metabolismo celular, atuando, por exemplo, nas redes reguladoras que controlam a expressão dos genes. Em organismos eucarióticos, como os seres humanos, os exons costumam compor apenas cerca de 10% de todo o material genético. Em organismos procarióticos há muito menos regiões não-codantes.

C. Aminoácidos

Em células eucarióticas, a informação presente no DNA é transformada em RNA que é passada para fora do núcleo da célula, onde as proteínas são finalmente sintetizadas. As células procarióticas (organismos mais simples) não possuem núcleo e a síntese de proteínas pode ocorrer imediatamente após a cópia da fita de DNA ou até durante esse processo. As proteínas, por sua vez, são compostas por pequenas sub-unidades presentes no citoplasma da célula chamadas aminoácidos. Uma seqüência de três letras de um DNA ou RNA corresponde a um códon, e cada códon é responsável por codificar um aminoácido em especial. Por exemplo, a seqüência de RNA:

AAGUCTTAGACU

Corresponde aos códons:

AAG UCT TAG ACU

Estes, por sua vez, especificam uma seqüência de aminoácidos. Existe na natureza um total de 20 aminoácidos diferentes (certos aminoácidos têm associados a si múltiplos códons) e seqüências com diferentes combinações destas moléculas formam os mais variados tipos de proteínas.

D. Proteínas

As proteínas são sintetizadas a partir da molécula de DNA e atuam nos processos metabólicos e estruturais de um organismo. Cada proteína tem a sua própria forma tri-dimensional e tipicamente possui de 1.000 a 50.000 átomos. Embora exista uma grande variação de estrutura e funcionalidade entre as proteínas, todas elas podem ser representadas por uma seqüência linear dos aminoácidos que as compõem. Esta seqüência é chamada de estrutura primária da molécula de proteína. Entretanto, a estrutura primária de uma proteína, em geral, não é suficiente para determinar sua forma tridimensional, a qual está intimamente relacionada com as suas propriedades e funções num organismo. É uma

tarefa extremamente difícil inferir com precisão a estrutura tri-dimensional de uma proteína baseado na sua seqüência primária. Esta é uma das questões mais estudadas em bioinformática (BALDI & BRUNAK, 2001).

1.2. Expressão Gênica

A. Transcrição e tradução

O processo de síntese de uma proteína ocorre no citoplasma da célula, enquanto as informações necessárias para construí-la se encontram no DNA. Para que o processo de síntese ocorra é necessário que haja uma transferência da informação presente no DNA para os ribossomos, estruturas responsáveis pela montagem da proteína através da concatenação de aminoácidos.

De forma simplificada, o transporte da informação codificada ocorre da seguinte maneira. Quando a célula necessita de uma determinada proteína, a informação presente no gene que codifica esta proteína deve ser copiada. As duas fitas do DNA são então separadas com a ajuda de enzimas especiais na região correspondente ao gene que está sendo solicitado. O conteúdo do gene é então copiado de forma complementar em uma fita de RNAm (RNA mensageiro). Este processo é conhecido como transcrição. O RNAm, por sua vez, se associa a um ribossomo presente no citoplasma. Os ribossomos são então responsáveis por interpretar a informação codificada em forma de RNA, associando as seqüências de três nucleotídeos (códons) aos seus aminoácidos correspondentes, ligando-os um a um e sintetizando a proteína. Esta etapa é chamada tradução.

Esse processo em forma de cadeia linear de síntese de proteínas a partir da informação dos genes é conhecido como *dogma central da biologia molecular*. A Figura 1.2 ilustra o sentido do fluxo de informação nesse processo.

Durante o processo de transcrição e tradução podem ocorrer várias etapas intermediárias, chamadas pós-transcrição e pós-tradução, em que o RNA e as proteínas são pré-processados antes de se tornarem efetivos. Essas etapas intermediárias não são destacadas na figura, mas o leitor interessado pode consultar ALBERTS *et al*. (1989).

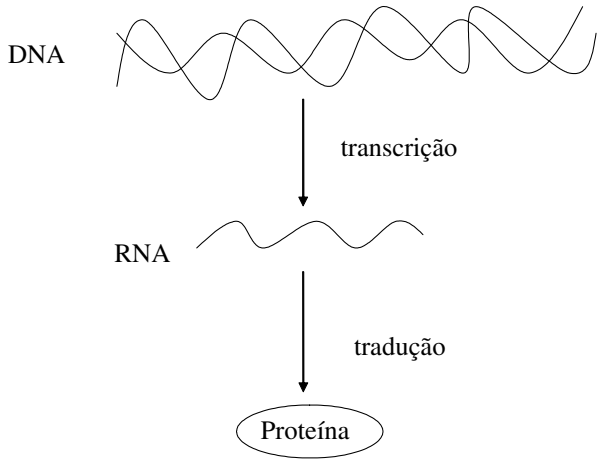


Figura 1.2 Dogma central da biologia molecular. A informação parte do DNA é transcrita em RNA e traduzida em proteínas.

Praticamente todas as células de um organismo multicelular possuem o genoma completo do indivíduo. Um fato intrigante é que, mesmo tendo em seu núcleo o mesmo material genético, células de diferentes órgãos possuem funções completamente distintas e as proteínas necessárias para desempenhar essas funções também são muito diferentes. Em casos como esses, em que houve diferenciação das células, os genes do DNA que se expressam não são os mesmos, sendo que um gene é considerado expresso toda vez que a proteína que ele codifica é sintetizada. Esse fenômeno ocorre também em uma mesma célula, pois durante o seu desenvolvimento ela vai necessitar de proteínas diferentes de acordo com estímulos internos ou externos, fazendo a expressão dos seus genes variar ao longo do tempo. A forma como os genes se comportam, isto é, quando eles devem ou não se expressar, é controlada pelas redes reguladoras, um mecanismo extremamente sofisticado capaz de interpretar os estímulos aos quais a célula está submetida, tais como a concentração de determinados elementos químicos, iniciando ou suprimindo a expressão.

Um fato de interesse nesse processo de ativação de um gene é que, como dito acima, toda vez que o gene é expresso ocorre a sua transcrição em forma de RNAm. Isso significa que o nível de expressão de todos os genes de um genoma são refletidos indiretamente nas concentrações de seus RNAm correspondentes. Essas concentrações, por sua vez, podem ser um forte indicador do estado biológico da célula, já que, em princípio, representam todas as proteínas que são sintetizadas pelos ribossomos. Esse é o princípio no qual se baseia o estudo da expressão de genes. Pode-se estudar os processos biológicos em um

organismo através da análise dos níveis de expressão de seus genes, que são obtidos através da leitura das concentrações de RNAm existentes em suas células.

B. Microarranjos de DNA: medindo a expressão gênica

O seqüenciamento de genomas completos de organismos criou uma forte base para estudos em genômica funcional. A determinação das seqüências, no entanto, embora seja uma fase fundamental para o estudo das funções dos genes, representa apenas uma pequena parte das possibilidades de análise. É possível também utilizar as informações do seqüenciamento em escala genômica para realizar estudos mais completos. Nesse sentido, diversas técnicas experimentais foram desenvolvidas, como *gene disruption* (ROSS-MACDONALD *et al*., 1999), *two-hybrid studies* (UERTZ *et al*., 2000), *large-scale proteomics* (CHRISTENDAT *et al*., 2000), *silicone elastomer protein chips* (ZHU *et al*., 2000), *serial analysis of gene expression* (SAGE) (VELCULESCU *et al*., 1997), e várias tecnologias de *microarrays* de DNA. Dessas técnicas, as de *microarrays* se tornaram particularmente populares devido ao alto paralelismo dos experimentos e à possibilidade de estabelecer relações estatísticas entre os dados obtidos (BERTONE & GERSTEIN, 2001).

Microarrays (ou microarranjos) de DNA são capazes de medir o nível de expressão de dezenas de milhares de genes simultaneamente, sob diferentes situações experimentais ou ao longo do tempo. Técnicas mais antigas já possuíam a habilidade de medir a expressão, mas o número de genes era bastante reduzido. O desenvolvimento dos *microarrays* permitiu uma revolução nos estudos em genômica, pois houve uma grande mudança quantitativa na escala dos experimentos, que levou a uma mudança qualitativa nas análises efetuadas, dando oportunidade para estudar o comportamento regulador dos processos biológicos em nível celular.

A habilidade de medir a expressão gênica traz a possibilidade de reduzir a dependência de conhecimentos prévios nas pesquisas, deixando para o conjunto de dados o papel de indicar direções promissoras nas investigações. Através da análise desses dados é possível determinar o papel funcional de vários genes, estudar a forma como os níveis de expressão refletem processos biológicos de interesse (como no caso de doenças), determinar os efeitos de tratamentos experimentais, além de permitir a criação de ferramentas para realizar diagnósticos baseados na regularidade dos padrões de expressão.

Um bom exemplo é o estudo feito por GOLUB *et al*. (1999), onde dois tipos de câncer, leucemia mielóide aguda e leucemia linfoblástica aguda, foram corretamente distinguidos através do estudo dos níveis de expressão gênica de tecidos cancerosos, sugerindo uma estratégia genérica para descobrir e predizer outros tipos de câncer.

1.3. Redes Reguladoras

A. Controle da expressão

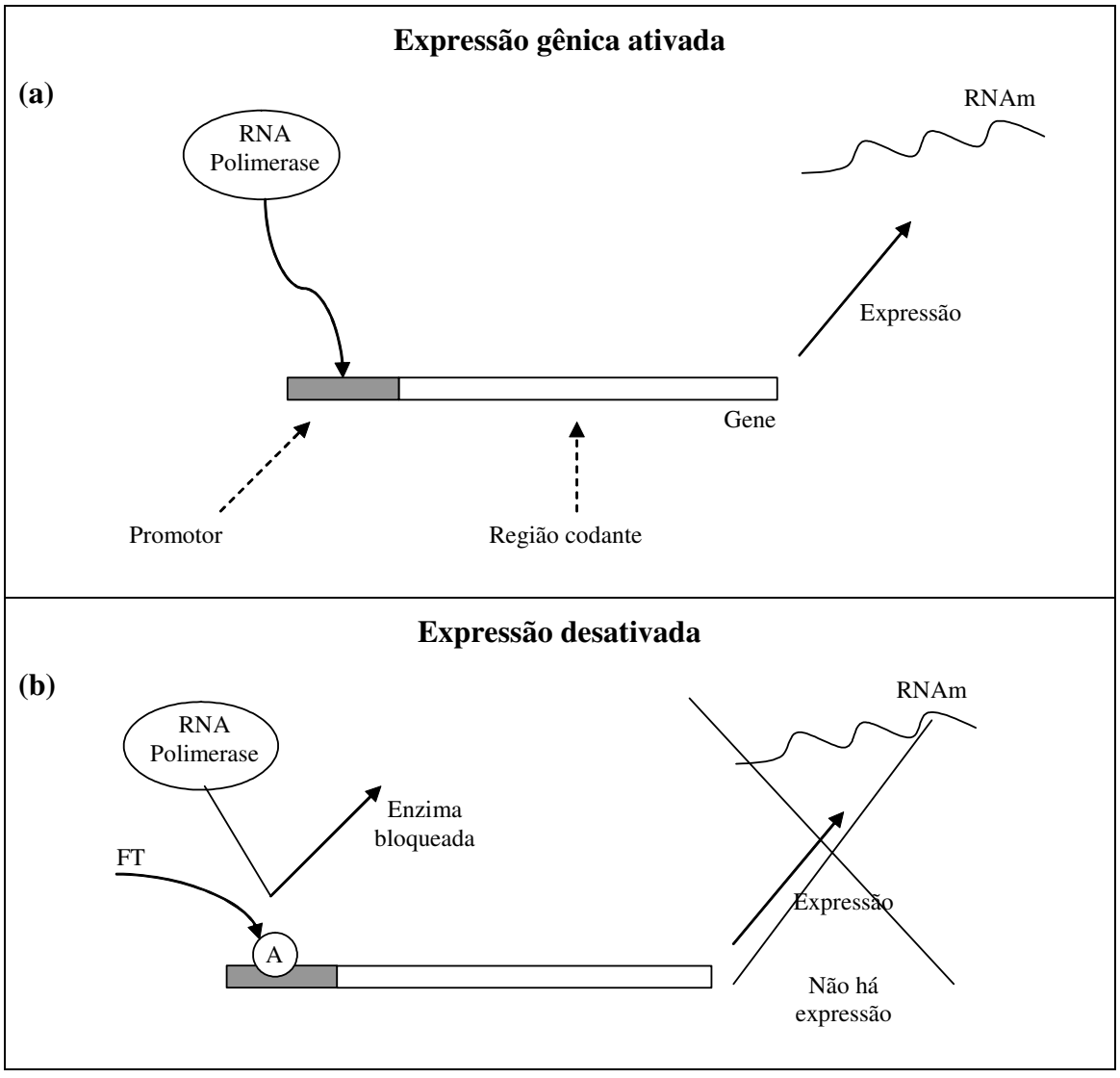
Até agora foi discutido, de forma simplificada, como se dá a expressão gênica e o papel desse processo para o funcionamento e diferenciação das células. Como dito anteriormente, as proteínas, que são produzidas pelos genes, são as unidades estruturais e funcionais das células. Porém, além de executar essas tarefas, uma grande parte das proteínas, conhecidas como fatores de transcrição (FT), são também capazes de realizar papéis reguladores, controlando a expressão dos genes. Essas proteínas determinam o momento em que um gene deve se expressar e a que taxa.

Para tentar compreender como funciona o processo de regulação, vamos olhar em mais detalhe como é feita a transcrição do material genético em RNA, isto é, a expressão gênica. A Figura 1.3 apresenta uma ilustração desse processo. Na Figura 1.3(a), o gene é dividido em duas partes: a região codante, que compreende à informação útil para a síntese de proteína e que é a parte do gene efetivamente transcrita para RNA, e o promotor, uma região que não é transcrita, mas é onde a enzima RNA-polimerase (a enzima que realiza a cópia da fita em RNA) deve se ligar primeiro para que a transcrição tenha início.

A figura esquematiza a enzima RNA-polimerase se ligando ao promotor e realizando a cópia do material genético em uma fita de RNA mensageiro. Na Figura 1.3(b), um fator de transcrição está presente (a proteína *A*) e ele se liga ao promotor do gene, inibindo a expressão por impossibilitar a enzima RNA-polimerase de iniciar a transcrição. Na Figura 1.3(c), outro fator de controle está presente. A proteína indutora *B* reage com a proteína *A*, formando o dímero *AB*. Por ter propriedades estruturais diferentes, esse dímero não pode se ligar ao promotor, e agora a expressão do gene é reativada.

O mecanismo contrário pode ocorrer, ou seja, a expressão do gene é originalmente desativada, pois o seu promotor não permite o acoplamento da RNA-polimerase. Mas um

fator de transcrição, ao se ligar ao promotor, pode mudar a conformação estrutural deste segmento de DNA, permitindo agora que a RNA-polimerase se ligue ao gene, iniciando sua transcrição.



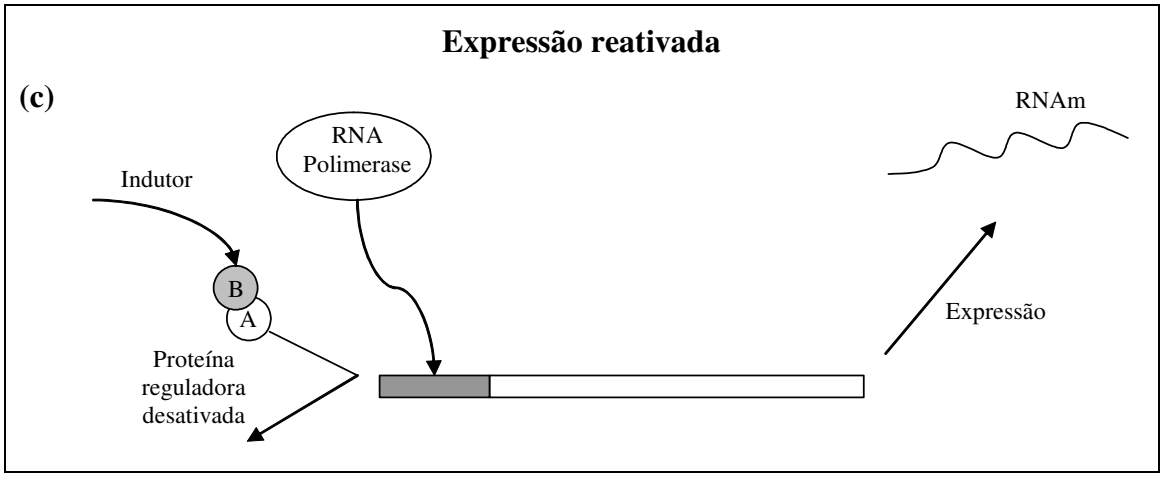


Figura 1.3 Regulação gênica através de fatores de transcrição. (a) Sem interferência de fatores de transcrição, a expressão do gene ocorre livremente. (b) Quando o fator de transcrição está presente, a expressão não ocorre mais. (c) A proteína indutora desativa o fator de transcrição, liberando novamente a transcrição.

Dois pontos devem ser salientados a respeito deste processo. Primeiro, a expressão do gene, geralmente, não é totalmente reprimida ou totalmente ativada. Cada fator de transcrição vai exercer uma influência diferente, aumentando ou diminuindo em diferentes graus a afinidade da RNA-polimerase pelo promotor. Segundo, a influência reguladora da proteína não varia apenas com o efeito que ela produz sobre a afinidade da RNA-polimerase com o promotor, mas também pela sua própria afinidade com o promotor e pela sua concentração.

Em termos moleculares, o cenário pode ser descrito da seguinte forma. Suponha que uma proteína reguladora inibe totalmente a transcrição quando está ligada ao promotor. Se esta proteína está presente, a RNA-polimerase não se liga ao gene, mas se ela está ausente, a enzima pode iniciar a cópia da fita. Mas a ligação da proteína com o promotor é uma reação bioquímica de dois sentidos (ida e volta), e como os eventos são probabilísticos, a proteína se liga, mas também se desliga do promotor. Portanto, havendo uma concentração constante de proteína reguladora, parte do tempo o promotor vai ficar livre e parte do tempo ocupado pela proteína, e esse tempo vai depender das constantes da reação, isto é, da afinidade da proteína com o promotor. No tempo em que ele está livre, a RNA-polimerase pode realizar a transcrição. Haverá, portanto, um tempo médio em que o promotor está livre e em que ele está ocupado, e, logo, uma expressão média diferente de zero, mesmo com a

proteína reguladora presente. Se a concentração da proteína reguladora aumenta, o que acontece é que a reação de ligação com o promotor é desequilibrada no sentido de ida. Portanto, o tempo médio em que o promotor está livre diminui, e a expressão é, conseqüentemente, reduzida também.

Considere agora mais algumas particularidades envolvidas no processo de regulação. A Figura 1.4 mostra um gene que pode receber a influência de mais de uma proteína reguladora. Na Figura 1.4(a) a relação entre as proteínas é cooperativa, pois elas podem se ligar simultaneamente ao promotor e cada combinação entre as proteínas vai gerar um efeito diferente sobre a expressão. Numa segunda situação (Figura 1.4(b)), a ligação com o promotor é competitiva. Embora várias proteínas possam reagir com ele, apenas uma proteína reguladora é permitida por vez.

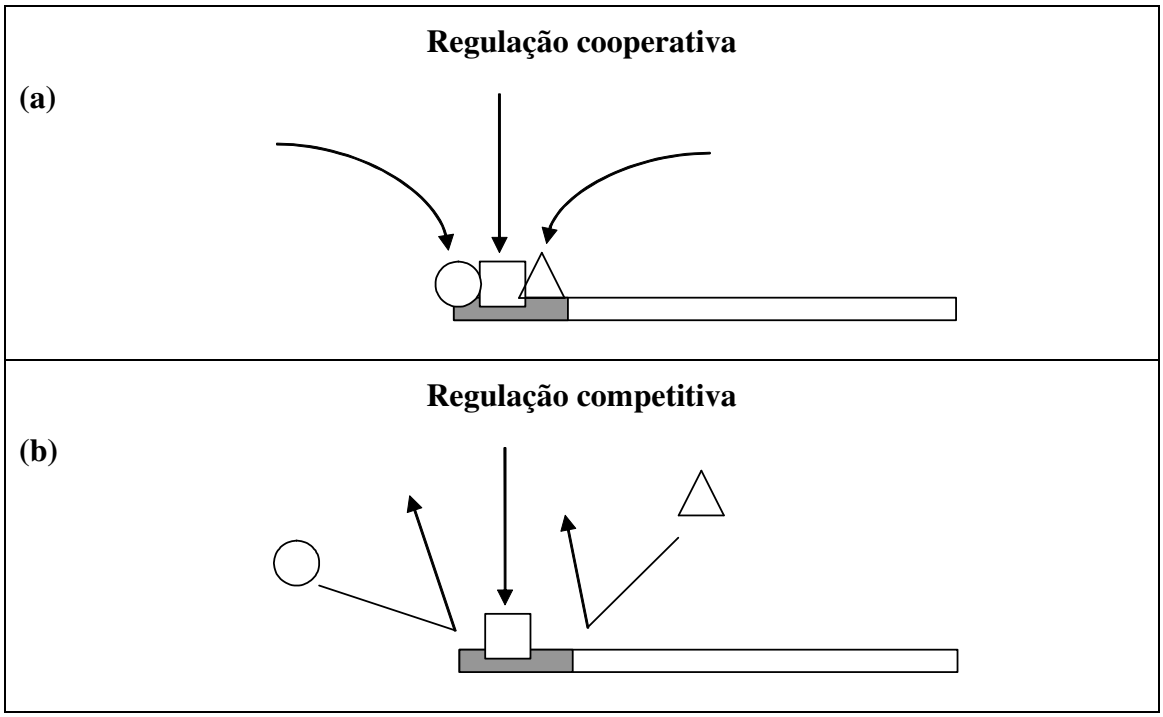


Figura 1.4 Tipo de regulação em que a expressão do gene é regulada por mais de um tipo de proteína. (a) Regulação cooperativa: as proteínas podem se ligar simultaneamente ao promotor. (b) Regulação competitiva: o promotor só permite a ligação de uma proteína por vez.

B. Controle em rede

Como descrito acima, um gene pode sofrer regulação através dos fatores de transcrição, e um mesmo gene pode ser regulado por várias proteínas diferentes. Além

disso, vimos que as variáveis envolvidas nesse processo são as constantes cinéticas das reações bioquímicas e as concentrações de cada molécula, e que a regulação pode implementar funcionalidades lógicas diferentes, isto é, cooperativa (OR) e competitiva (AND).

Todas essas considerações foram realizadas analisando-se apenas um gene. No entanto, deve-se ter em mente que cada proteína reguladora é produzida por um gene também, e que este gene, por sua vez, é regulado por proteínas reguladoras produzidas por outros genes. Além disso, as próprias proteínas reguladoras reagem entre si, determinando seus estados de ativação ou inativação. Como resultado, o controle da expressão é realizado por uma rede de interações gênicas e protéicas, a chamada rede reguladora (ou rede gênica).

A Figura 1.5 dá uma ilustração das implicações de uma rede reguladora muito simples, com apenas 3 genes.

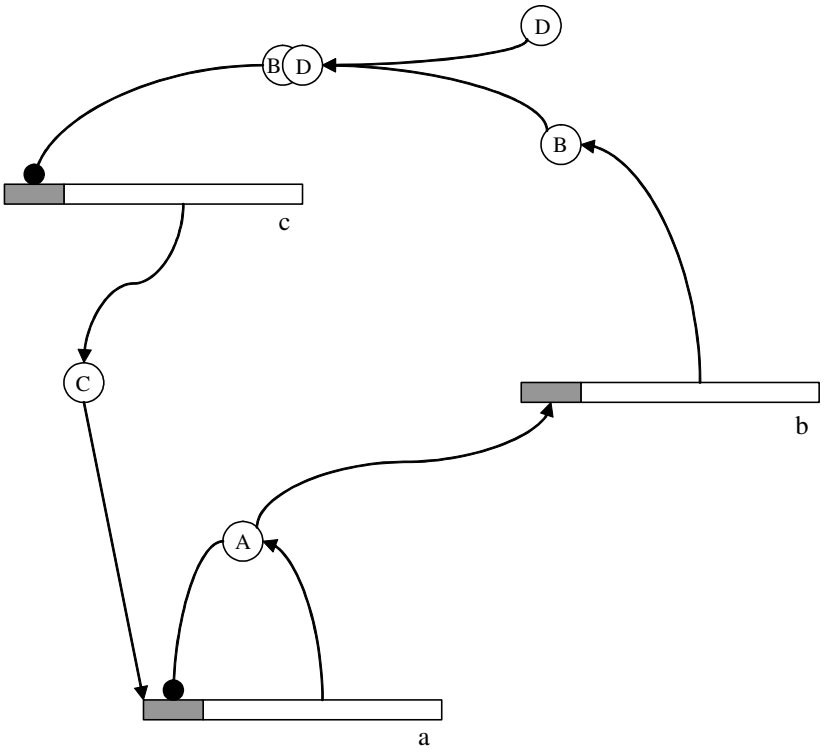


Figura 1.5 Ilustração de uma rede reguladora com apenas 3 genes. Setas indicam interação estimulatória e círculos em preto, interação inibitória.

Nesse esquema, o gene *a* produz a proteína *A* que inibe a sua própria produção. A proteína *A* também regula o gene *b*, estimulando a produção de *B*. A proteína *B*, por sua vez, quando se liga com um fator externo *D*, forma um complexo ativo *BD*. Esse complexo

inibe a expressão do gene *c*. A proteína *C*, produzida pelo gene *c*, estimula a produção de *A*. Note que o processo intermediário de produção de RNA não está sendo modelado aqui (a aproximação supõe que a expressão gênica corresponde diretamente à síntese de proteínas), mas ele pode ser inserido de forma a obter uma representação mais realista.

Como resultado desse esquema, temos um sistema dinâmico acoplado bastante complexo, regido por eventos probabilísticos e onde a concentração das proteínas varia o tempo todo. É possível supor o que acontece com a complexidade desse sistema quando o número de variáveis aumenta para a ordem de milhares.

1.4. Modelagem Computacional das Redes Reguladoras

Regida por equações não-lineares estocásticas e circuitos de realimentação positiva e negativa, a dinâmica das redes gênicas é muito complexa quando o número de variáveis envolvidas é grande. Torna-se muito difícil, neste caso, obter uma compreensão intuitiva do funcionamento dessas redes. Neste cenário, o uso de técnicas de modelagem e simulação computacional se torna fundamental para o estudo desses sistemas.

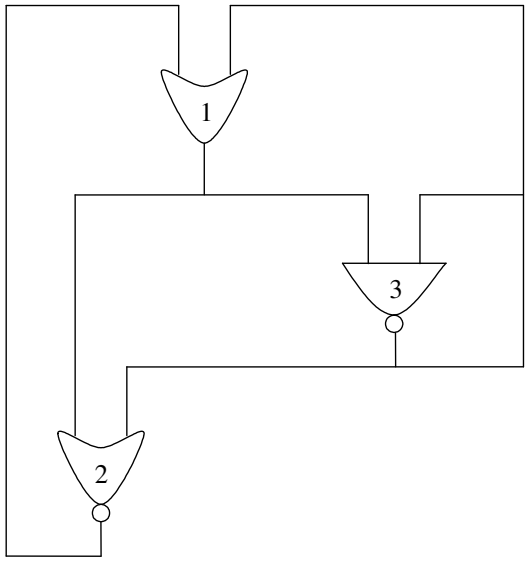
Nesta seção, as principais metodologias de modelagem das redes gênicas utilizadas na literatura são apresentadas e descritas brevemente.

A. Redes booleanas

As redes booleanas são baseadas em uma simplificação grosseira do funcionamento dos mecanismos reguladores. A hipótese adotada é que um gene tem apenas dois estados discretos possíveis, ativo e inativo, e, com isso, é possível empregar uma modelagem dessas redes baseada em uma lógica booleana. Em outras palavras, uma rede gênica corresponde a um circuito lógico em que cada gene pode assumir valor 1 (ativo) ou 0 (inativo). Como esse circuito é realimentado, as redes booleanas se tornam um sistema dinâmico e o estado dos genes é atualizado discretamente a cada iteração.

Considerando um sistema com n variáveis (genes) x_i, $1 \leq i \leq n$, temos que o espaço de estados do sistema tem 2^n possíveis valores diferentes. O estado de cada variável no próximo instante de tempo t+1 é então determinado pelas entradas da sua função lógica no instante atual t. Se para cada função booleana tivermos k entradas, o número total de

funções boolenas possíveis será de 2^{2^k}. A Figura 1.6(a) apresenta uma ilustração de uma rede booleana com 3 variáveis e com duas entradas possíveis para cada uma delas. Na Figura 1.6(b), vemos como o estado das variáveis são atualizados através das equações definidas pelas funções lógicas.



$$x_1(t+1) = x_2(t)\ or\ x_3(t)$$
$$x_2(t+1) = x_1(t)\ nor\ x_3(t)$$
$$x_3(t+1) = x_1(t)\ nand\ x_3(t)$$

(a) (b)

Figura 1.6 (a) Rede booleana com 3 variáveis e duas entradas por função lógica. (b) Equações de atualização dos estados das variáveis para a mesma rede.

Dada sua concepção simplificada, as redes booleanas são adequadas para simular redes gênicas em grande-escala. Na literatura, elas têm sido utilizadas para estudar as propriedades globais de sistemas reguladores (KAUFFMAN, 1993; SOMOGYI & SNIEGOSKY, 1996; SZALLASI & LIANG, 1998; WEISBUCH, 1986). A idéia básica é gerar redes booleanas com propriedades locais de interesse, como, por exemplo, diferentes números de outros genes reguladores (o parâmetro *k* definido acima) ou diferentes tipos de funções booleanas, e avaliar a influência desses fatores na regulação gênica. Localizando atratores, trajetórias do sistema e bacias de atração no espaço de estados, é possível investigar sistematicamente as implicações das propriedades locais para a dinâmica global das redes.

Como exemplo dessa aplicação é possível citar o trabalho de KAUFFMAN (1993). Utilizando redes booleanas aleatórias de até 10.000 variáveis, Kauffman mostrou que, para valores pequenos de *k* e com funções booleanas escolhidas também aleatoriamente, o sistema exibe dinâmica bastante ordenada. Para essas redes, foi mostrado empiricamente

que o número de atratores médio esperado é de $\sqrt{n}$ e que o período dos atratores periódicos encontrados (ciclos limite) é também proporcional a $\sqrt{n}$.

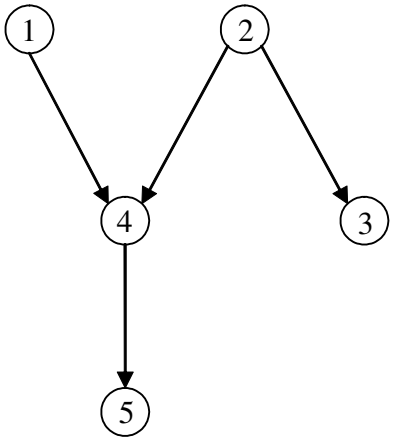
Redes booleanas são uma opção em que a especificidade e o realismo do sistema são abdicados em troca do estudo de propriedades globais. É uma abordagem válida quando considerada em conjunto com propostas mais realistas.

B. Redes bayesianas

As redes bayesianas são um método estatístico formal (HECKERMAN, 1997) para descrever um sistema estocástico através de relações causais. Uma rede bayesiana pode ser representada por um grafo acíclico $G = \{V,A\}$, como ilustrado na Figura 1.7. Os vértices $i \in V$, $1 \leq i \leq n$, representam as variáveis do sistema, que são variáveis aleatórias. Na modelagem de redes gênicas, as variáveis correspondem aos genes e as arestas do grafo às interações reguladoras entre eles.

Numa rede bayesiana, o estado de cada variável é determinado por uma função de densidade de probabilidade condicional, em que a probabilidade de um gene assumir um determinado valor depende das funções de densidade de probabilidade dos genes pais. Os genes pais de uma variável são todas aquelas variáveis que possuem um arco dirigido à variável filha, ou seja, são os reguladores diretos de um gene. Esse conjunto de variáveis pais mais variáveis filhas é chamado família. Formalmente, a distribuição condicional de cada variável X_i é igual a $p(X_i \mid \text{pais}(X_i))$. Portanto, para a rede bayesiana da Figura 1.7(a), as probabilidades de cada variável são determinadas pelas relações de dependência das famílias. Essas relações são mostradas na Figura 1.7(b), juntamente com a densidade de probabilidade da rede como um todo, $p(X)$.

Veja que, para o cálculo da probabilidade condicional de cada variável, as variáveis pais são tomadas como independentes, mesmo que elas sejam na verdade dependentes. Essa independência condicional é chamada *independência de Markov* e facilita muito o cálculo das probabilidades, pois cada família pode ser considerada isoladamente, e a função de densidade de probabilidade do modelo é depois calculada através do produtório das probabilidades das famílias.



$$p(X_1)$$
$$p(X_2)$$
$$p(X_3 \mid X_2)$$
$$p(X_4 \mid X_1, X_2)$$
$$p(X_5 \mid X_4)$$

$$p(X) = p(X_5 \mid X_4)p(X_4 \mid X_1, X_2)p(X_3 \mid X_2)p(X_2)p(X_1)$$

(a) (b)

Figura 1.7 (a) Estrutura de uma rede bayesiana. (b) Probabilidades condicionais para cada uma das variáveis da rede e a função de densidade de probabilidade da rede inteira.

As redes bayesianas são muito utilizadas como método de inferência para determinar as relações reguladoras a partir de dados de expressão gênica. Isto é, dado um conjunto de dados na forma de variáveis independentes X_i, é possível realizar uma busca no espaço de todas as possíveis estruturas de redes bayesianas de forma a encontrar a rede que melhor explica as amostras disponíveis (baseado na maximização de um critério de qualidade). Como essa otimização é um problema do tipo *NP*-difícil (CHICKERING *et al*., 2004), métodos heurísticos de busca combinatória são geralmente necessários.

Utilizando aprendizado em redes bayesianas, PE'ER *et al*. (2001) estudaram as relações de regulação dos genes envolvidos no ciclo de vida celular da levedura do pão *S. cerevisiae*. Os dados de expressão originais continham 6.177 genes e 76 condições experimentais, e o algoritmo de inferência de redes foi aplicado a 800 genes cujos valores de expressão variaram mais significativamente. Analisando as interações evidenciadas pelo algoritmo, foi mostrado que apenas alguns poucos genes dominam o processo de regulação que dá origem ao ciclo celular. Muitos desses genes são de fato conhecidos como estando envolvidos no controle e iniciação do ciclo celular.

As redes bayesianas são uma ferramenta muito interessante para a análise dos dados de expressão gênica, pois permitem a investigação da estrutura de relacionamento dos genes, representando, portanto, uma oportunidade para mapear as redes reguladoras. Além disso, essas redes possuem caráter probabilístico (não-determinístico) o que é mais coerente

com o funcionamento dos sistemas reais. No entanto, as redes bayesianas são geralmente estáticas, e essas estruturas não condizem com a natureza dos sistemas reguladores. Essa limitação pode ser contornada com a utilização de modelos generalizados, como as redes bayesianas dinâmicas (FRIEDMAN *et al*., 1998).

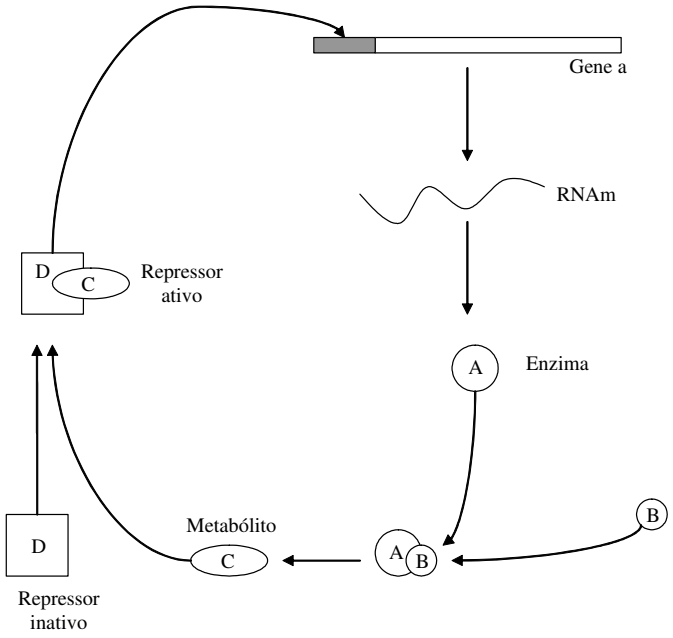
C. Equações diferenciais

A modelagem de redes gênicas por equações diferenciais ordinárias é, possivelmente, a metodologia mais amplamente utilizada para representar e simular as redes no computador. Neste formalismo, as concentrações de RNAs, proteínas e outras moléculas são modeladas como variáveis no tempo assumindo valores reais não negativos. As interações reguladoras tomam a forma de relações funcionais e diferenciais entre as concentrações das variáveis.

Mais especificamente, as relações entre as variáveis são modeladas por equações de taxa de produção (*rate equations*), um método popularmente utilizado em cinética química, em que as reações químicas são descritas como equações diferenciais acopladas, expressando a taxa de produção (aumento de concentração) de uma variável em função (da concentração) de outras.

Considere o exemplo de sistema regulador da Figura 1.8, adaptado de (GOODWIN, 1963). A figura mostra um sistema regulador simples de apenas um gene, considerando a produção de RNA. As equações da Figura 1.8(b) descrevem o comportamento de algumas variáveis do sistema. A função *r* pode ser representada pela função sigmoidal *Hill curve*, mostrada na Figura 1.9. Essa função será definida formalmente no Capítulo 2.

Os modelos de equação diferencial têm sido utilizados para estudar circuitos genéticos pequenos. A maior parte dos estudos analisa o papel dos circuitos de realimentação positiva e negativa (CHERRY & ADLER, 2000; GOODWIN, 1965; KELLER, 1994; SMOLEN *et al*., 2000). Realimentação negativa tem sido associada a comportamentos oscilatórios, muito importantes para o metabolismo celular (veja o Capítulo 4). Já a realimentação positiva está associada à possibilidade de múltiplos estados estacionários. De fato, a instabilidade ocasionada pela realimentação positiva aliada à saturação é responsável por produzir mais de um estado estável e essa multi-estacionaridade tem sido associada aos estados de diferenciação celular (THOMAS, 1998).



$$\dot{X}_{RNAm} = k_{RNAm} r(X_C) - \gamma_{RNAm} X_{RNAm}$$
$$\dot{X}_A = k_A X_{RNAm} - \gamma_A X_A$$
$$\dot{X}_C = k_C X_A - \gamma_C X_C$$

(a) (b)

Figura 1.8 (a) Sistema regulador envolvendo a síntese de *RNAm*, a produção de uma enzima *A*, a reação enzimática de *A* com o substrato *B*, produzindo o metabólito *C*, a ativação do repressor *D* através de *C* e a regulação do gene *a*. (b) As equações diferenciais que modelam o comportamento das concentrações *RNAm*, *A* e *C* são mostradas, onde *X* representa a concentração de cada molécula, *k* as constantes cinéticas de produção e γ as constantes de degradação. A função *r* representa uma curva de regulação não-linear variando de zero a um.

As equações diferenciais têm sido utilizadas com sucesso na modelagem de diversos circuitos conhecidos (BORISUK & TYSON, 1998; HAMMOND, 1993; MACADAMS & SHAPIRO, 1995; MAHAFFY, 1984; REINITZ & VAISNYS, 1990), e a simulação em computador desses sistemas têm ajudado a desenvolver uma noção intuitiva do comportamento regulador. Atualmente, a principal dificuldade com essa técnica de modelagem é a ausência de informações específicas sobre as constantes cinéticas. Geralmente esses parâmetros são determinados em experimentos *in vitro*, mas *in vivo*, devido à interferência de fatores celulares internos, as constantes assumem valores bastantes diferentes e os modelos computacionais muitas vezes acabam não representando bem os fenômenos observados.

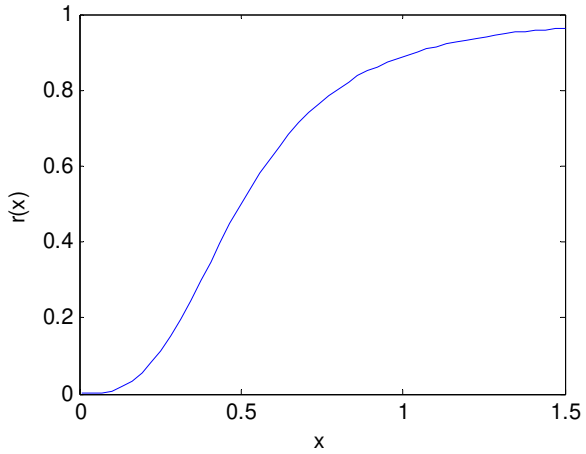


Figura 1.9 *Hill curve* para parâmetros arbitrários, onde *x* corresponde à ação reguladora e *r* é o valor retornado pela função.

D. Equações estocásticas

A modelagem por equações diferencias pressupõe que as concentrações das substâncias variam continuamente e deterministicamente, duas suposições que podem ser questionadas no caso de regulação gênica (GIBSON & MJOLSNESS, 2001; GILLESPIE, 1977). Em primeiro lugar, em algumas situações o número de moléculas envolvidas num processo regulador é muito pequeno (da ordem de dezenas), o que compromete a suposição da modelagem contínua. Segundo, as mudanças determinísticas pressupostas pelas equações diferenciais podem ser questionáveis devido a flutuações de tempo nos eventos celulares, como no atraso entre o início e o fim da transcrição. Como conseqüência, dois sistemas reguladores iguais com as mesmas condições iniciais podem acabar se encaminhando para estados diferentes, um fenômeno que é agravado quando o número de moléculas envolvidas é reduzido.

Para tentar contornar essas limitações, alguns autores propuseram modelos discretos e estocásticos da regulação gênica (GILLESPIE, 1977; ARKIN *et al*., 1998). Nesses modelos, quantidades discretas de moléculas são as variáveis de estado do sistema, e uma distribuição de probabilidade conjunta é introduzida para expressar a probabilidade de que, em um dado instante, a célula assuma um determinado estado.

A simulação dessas equações é geralmente realizada por meio de um método chamado simulação estocástica (*stochastic simulation*), proposto por GILLESPIE (1977).

Basicamente, o algoritmo de simulação estocástica determina quando a próxima reação ocorre (através da probabilidade de encontro entre moléculas) e de que tipo ela será, dado o estado do sistema. Em seguida, o estado do sistema é atualizado e o processo se inicia novamente.

A simulação estocástica foi utilizada por MCADAMS & ARKIN (1997) para analisar as interações que controlam a expressão de um único gene procariótico. Eles investigaram como o intervalo de tempo entre a ativação de um gene e a ação reguladora do seu produto em outro gene, o chamado tempo de comutação, é afetado pela natureza estocástica dos intervalos de transcrição e do número de moléculas produzidas. Eles mostraram que, para este gene, rajadas de transcrição são produzidas em intervalos de tempo aleatórios, levando a grandes flutuações no tempo de comutação.

Os resultados da simulação estocástica estão mais próximos da realidade da regulação gênica, mas o uso dessa técnica nem sempre é evidente. Em primeiro lugar, a abordagem requer conhecimento detalhado dos mecanismos das reações envolvidas, incluindo as funções de densidade de probabilidade. Além disso, as simulações são geralmente muito custosas em relação a outras técnicas de modelagem, o que limita a sua aplicação.

E. Matriz de pesos

Uma matriz de pesos (WEAVER *et al*., 1999) consiste numa matriz $n \times n$, onde n é o número de genes, e os pesos (elementos da matriz) indicam a influência reguladora de um gene sobre outro. Os pesos W_{ij} representam a influência do gene i sobre o gene j, e a entrada reguladora total para um dado gene j é dada pela soma de todas as entradas i, multiplicadas pelos seus respectivos pesos. A matriz de pesos considera a interação de todas as combinações de genes, muitas das quais terão peso zero. Após a somatória da entrada, a saída da expressão do gene é determinada por uma função sigmoidal, provendo não-linearidade ao modelo.

Essa estrutura de entrada-saída corresponde à estrutura de uma rede neural realimentada. Os pesos da matriz são inicialmente desconhecidos, mas podem ser determinados de forma a se obter uma dinâmica desejada utilizando meta-heurísticas de otimização, como *simulated annealing* ou algoritmos genéticos. Essas matrizes foram

utilizadas por REIJITZ & SHARP (1995) para modelar o comportamento do gene *eve* da mosca *Drosophila melanogaster*.

1.5. Estrutura das Redes Gênicas e Protéicas

Foi dito que o mecanismo de regulação assume a forma de uma rede de interações gênicas e protéicas. Veremos agora que essa rede possui uma estrutura organizada e que as propriedades estruturais da rede podem ter implicações no funcionamento do sistema.

A. Estrutura em lei da potência

O mapeamento das redes celulares revelou que a estrutura dessas redes segue a chamada lei da potência (JEONG *et al*., 2000), ou seja, a probabilidade de um determinado nó ter k conexões é $p(k) = k^{-\lambda}$, onde λ é o fator de decaimento. Em outras palavras, a lei da potência indica que há uma grande quantidade de nós com muito poucas conexões e uma pequeníssima quantidade de nós com muitas conexões.

Juntamente com o mapeamento das redes celulares, mapas estruturais de vários outros sistemas complexos, como a internet, as redes neurais, as redes sociais e as redes de interações de espécies (BARABÁSI, 2002; SONG *et al*., 2005), começaram a ser disponibilizados na literatura. Uma análise comparativa desses mapas mostrou que todas essas estruturas também seguem a lei da potência, embora cada uma possua um fator de decaimento específico.

Essa descoberta causou bastante entusiasmo na comunidade científica, uma vez que vários sistemas, em princípio não relacionados, agora apresentavam uma forte ligação em termos de similaridade de organização estrutural. Essa nova visão levanta duas perspectivas. Primeiro, sugere que o princípio organizacional e de processamento de informação é potencialmente o mesmo para todos os sistemas complexos auto-organizados, particularmente para os sistemas vivos. Segundo, que as propriedades estruturais desses sistemas estão de fato relacionadas às suas propriedades funcionais, e que a estrutura deve ser considerada como um fator a ser analisado em conjunto com outras propriedades do sistema.

Logo em seguida à descoberta da estrutura em lei da potência das redes gênicas, diversas iniciativas apresentaram explicações semelhantes para a sua origem (HALLINAN, 2004). Segundo essas propostas, o primeiro ponto a ser considerado é que as redes não são estáticas, elas crescem e, no caso das redes gênicas, o processo evolutivo determina esse crescimento. Como segundo fator, foi mostrado através de simulações computacionais que o crescimento da rede gênica, quando é realizado através da duplicação de genes (ou seja, a cópia "acidental" de um gene em outra região do DNA durante o processo de reprodução, um fenômeno já bem conhecido em biologia (ALBERTS *et al*., 1989), é capaz de gerar uma estrutura em lei da potência, tanto em termos de interação gênica quanto em interação protéica, uma vez que o novo gene herdará características do gene pai, e as proteínas produzidas herdarão as interações.

B. Propriedades

Como conseqüência da estrutura em lei da potência, duas características interessantes emergem nessas redes. A primeira delas é que a média dos caminhos mínimos entre todos os nós (uma medida chamada *comprimento característico* ou *diâmetro* da rede) é muito pequena em relação ao número de nós, considerando o que se poderia esperar de uma rede aleatória. Caminho mínimo significa o menor número de arcos que se deve percorrer num grafo para chegar de um nó a outro. Essa propriedade de diâmetro pequeno em relação ao número de nós da rede é chama de *mundo pequeno* (do inglês *small world*) (WATTS, 1999). Numa rede de interação de proteínas com 6.000 a 7.000 nós, por exemplo, o diâmetro é de aproximadamente 3. Isso implica que, numa rede desse tipo, as informações em um extremo da rede podem se dispersar e influenciar todo o sistema rapidamente.

A segunda característica que emerge da estrutura em lei da potência é que, dado que a distribuição da conectividade não é igualitária, haverá alguns poucos nós com muitas conexões. Esses nós muito conectados são chamados *hubs*. ALBERT *et al*. (2000) mostram que uma estrutura em lei da potência é muito mais tolerante a falhas do que uma rede aleatória ou uma rede exponencial (ambas não possuem *hubs*). Nesse estudo, as falhas são consideradas como remoção de nós aleatórios da rede, e o dano causado pela falha é representado pelo aumento no diâmetro da rede. Uma rede do tipo lei da potência aumenta

mais lentamente em diâmetro com o aumento do percentual de falhas do que os outros tipos de rede. Isso acontece porque os *hubs* são a principal via de interligação entre os nós da rede e, como eles são muito menos numerosos, falhas aleatórias dificilmente serão capazes de afetá-los significativamente. No entanto, o estudo mostra que essas redes são extremamente vulneráveis a ataques inteligentes. Se apenas os nós mais conectados na rede são removidos, sua estrutura se desintegra rapidamente. De fato, experimentos com organismos reais demonstraram que a remoção das proteínas mais conectadas de uma rede celular geralmente causam a morte do organismo, enquanto a eliminação das proteínas menos conectadas não costuma ser letal (JEONG *et al*., 2001).

Outro papel importante dos *hubs* está relacionado à dispersão de informação. Por estar muito conectado, um *hub* fica muito suscetível a informações provenientes de outros nós. Caso o *hub* seja realmente influenciado, essa informação pode se dispersar rapidamente pela rede e mudar todo o comportamento do sistema. Como exemplo, é possível citar a dispersão de doenças em uma rede social. Uma pessoa com muitos contatos sociais se torna mais suscetível a entrar em contato com pessoas doentes e, portanto, de contrair uma doença contagiosa qualquer. Uma vez que essa pessoa é contaminada, ela poderá dispersar essa doença contagiosa muito mais rapidamente na população (WATTS, 1999).

C. Hierarquia modularizada

Um debate que causou certa polêmica na linha das redes gênicas é a existência ou não de módulos funcionais, isto é, um grupo de genes (ou proteínas) que em conjunto realizam uma operação específica. De fato, há evidências de grupos de proteínas especializadas em determinadas funções, mas essas evidências não são suficientes para concluir que a rede gênica é constituída por tais módulos.

Ao analisar a estrutura das redes intracelulares de 43 organismos, RAVASZ *et al*. (2002) observaram que todas essas redes possuíam estruturas muito semelhantes, exibindo fator de decaimento e também coeficiente de clusterização (uma outra medida quantitativa da estrutura) quase iguais. Tentando reproduzir essas estruturas em computador, eles mostraram que uma estrutura simplesmente do tipo lei da potência, mas sem módulos, é capaz de assumir o mesmo fator de decaimento, mas possui coeficiente de clusterização

diferente dos observados. Com uma estrutura apenas modular, a situação se inverte. O coeficiente de clusterização coincide com o das redes reais, mas o fator de decaimento se torna diferente. Esse dilema foi resolvido com a proposta de uma estrutura hierárquica modular, na qual módulos maiores são constituídos de módulos menores que, por sua vez, são constituídos de módulos menores ainda, e assim sucessivamente. Essa estrutura possui uma característica chamada auto-similaridade, por repetir os mesmos padrões em vários níveis hierárquicos e é, portanto, conhecida na literatura como estrutura fractal (RAVASZ *et al*., 2002). O padrão de estrutura fractal apresenta os mesmos coeficientes das redes reais, e, de fato, a existência de módulos dentro de outros módulos têm sido confirmada por evidências experimentais. Essa hipótese também está de acordo com evidências encontradas em outras redes, como no caso das redes neurais do cérebro, que constituem módulos especializados em determinadas funções, e dentro desses módulos há regiões menores correspondendo a tarefas mais específicas. O mesmo pode ser encontrado nas redes ecológicas, em que há nichos e sub-nichos de espécies.

A Figura 1.10 apresenta uma ilustração de uma rede hierárquica modular. Veja como a estrutura se modifica à medida que os níveis hierárquicos aumentam. A estrutura da figura é totalmente simétrica em todos os níveis. Obviamente, em organismos reais essa simetria perfeita não é esperada.

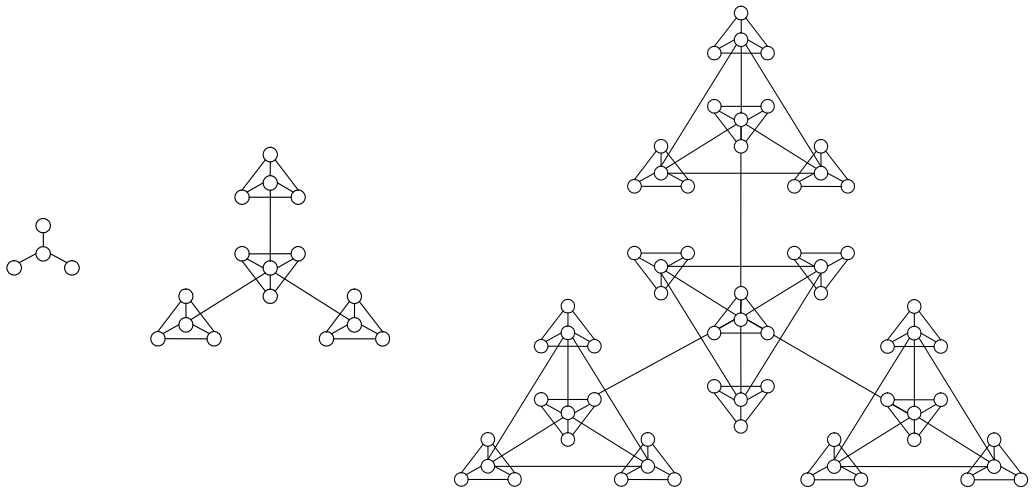


Figura 1.10 Formação de uma rede hierárquica modular em três níveis.

Capítulo 2

Recuperação de Redes Gênicas

Resumo – Este capítulo trata do problema de inferência de redes gênicas a partir de dados de microarranjos, utilizando redes bayesianas. A tarefa consiste em, a partir dos dados, gerar um modelo de rede bayesiana que explica o comportamento das variáveis (isto é, os níveis de expressão gênica observados) ao longo dos experimentos. Atualmente, a principal dificuldade relacionada a este problema é a ausência de amostragens suficientemente representativas para que a correlação entre as variáveis seja estimada com confiabilidade. A quantidade de dados disponível é geralmente muito reduzida considerando a complexidade da tarefa de inferência, e a situação é ainda agravada pelos níveis elevados de ruído dos dados de expressão. Levando isto em consideração, é proposto aqui um método de estimação de densidade de probabilidade que busca maximizar a utilização dos dados disponíveis, gerando representações aceitáveis em circunstâncias nas quais métodos tradicionais não operam satisfatoriamente. Este novo método é usado para capturar a correlação entre os genes na tarefa de inferência de redes bayesianas em domínio contínuo. A técnica proposta é comparada com uma metodologia de redes bayesianas discretas tradicionalmente aplicada a este problema.

2.1. Introdução

O problema de recuperação de redes gênicas consiste em, a partir de um conjunto de dados descrevendo o estado dos genes em circunstâncias diferentes, tentar inferir qual as relações causais determinantes para o comportamento observado do sistema. Em termos mais gerais, a tarefa consiste em gerar um modelo probabilístico que explique com o máximo de satisfação possível (segundo algum critério objetivo) um conjunto de dados observados, e esse modelo deve ser descrito na forma de relações causais. A proposição de modelos a partir de dados observados é denominada identificação de sistemas (AGUIRRE, 2004).

Para o caso específico de redes gênicas, os dados que descrevem os estados do sistema são os dados de expressão gênica, ou seja, o estado de cada gene pode ser representado pelo nível de expressão que ele apresenta em determinada circunstância. Assim, de forma a tentar recuperar a estrutura de interações genéticas de uma rede, uma prática comum é perturbar o sistema, expondo-o a diferentes condições experimentais, e medir os níveis de expressão gênica que são obtidos em resposta. Quanto mais abrangentes forem as condições experimentais, melhores serão as perspectivas de se chegar ao mapeamento das interações. Quando o interesse da análise reside na dinâmica do processo, os dados devem assumir características temporais. O modelo deve descrever as relações causais que explicam a dinâmica das variações observadas no sistema.

Modelos de redes gênicas que procuram explicar os comportamentos observados em sistemas celulares são de fundamental importância para o entendimento dos processos biológicos. A simples caracterização dos genes e de seus papéis, em geral, não é suficiente para explicar eventos e fenômenos celulares de interesse, simplesmente porque na grande maioria dos casos não há uma função específica para cada gene. Espera-se, porém, que através da análise de mapas de redes de interações seja possível descrever e compreender os processos em cadeia responsáveis por determinados estados fenotípicos. Um exemplo clássico é o mapeamento das interações que dão origem a uma doença como o câncer. Baseado na via de relações causais, seria possível, por exemplo, avaliar a viabilidade e eficácia de uma intervenção artificial em algum dos níveis intermediários visando interromper o processo.

Devido à importância que lhes é atribuída, a demanda por esses mapeamentos tem sido muito grande ultimamente. Não obstante, existe ainda uma carência de ferramentas computacionais capazes de gerá-los de forma sistemática, e uma das principais razões para isso é a falta de informação suficiente. Mais especificamente, os experimentos típicos de microarranjos, envolvendo cerca de 100 condições experimentais, não fornecem amostras com quantidade e representatividade suficientes para investigar com confiabilidade a correlação entre as variáveis, e, por conseguinte, produzir mapeamentos adequados. Como resultado, é necessário desenvolver métodos de inferência capazes de lidar com uma menor quantidade de informação, otimizando assim a utilização dos dados disponíveis e extraindo deles o máximo de conhecimento possível.

Tendo essas circunstâncias como motivação, é proposto nesse capítulo um método de estimação de densidade projetado especialmente para conjuntos de dados pequenos. Esse modelo é empregado num contexto de inferência de redes bayesianas para compor uma técnica de recuperação de redes gênicas que tende a maximizar a utilização dos dados de expressão.

Este capítulo está dividido, no que segue, em 4 seções. Na Seção 2.2, é apresentada uma contextualização da literatura, levantando uma discussão sobre as questões que motivaram o uso das redes bayesianas como técnica de modelagem e sobre a problemática envolvida no emprego das técnicas de estimação de densidade mais comumente utilizadas no processo de inferência. Na Seção 2.3 o algoritmo de estimação de densidade proposto é apresentado e alguns experimentos de estimação de densidade são realizados. A Seção 2.4 apresenta a metodologia a ser utilizada nos experimentos de inferência de redes gênicas estáticas e os resultados dos experimentos. A Seção 2.5 faz uma discussão sobre os resultados obtidos.

Antes de prosseguir com a leitura do capítulo, é sugerido ao leitor interessado que consulte o Apêndice, onde uma série de experimentos computacionais são realizados utilizando redes bayesianas. Os experimentos exploram o potencial das redes bayesianas como ferramentas de identificação de sistemas, e investigam a aplicabilidade prática das redes bayesianas a problemas do mundo real.

2.2. Aspectos Preliminares

Inferir redes gênicas confiáveis a partir de dados de expressão é uma tarefa bastante desafiadora. Algumas das principais dificuldades provêm da natureza dos processos genéticos em si, uma vez que as interações gênicas reguladoras são essencialmente não-lineares e os mecanismos de controle celulares, robustos a pequenas perturbações, são inerentemente estocásticos (MCADAMS & ARKIN, 1997). Ademais, devido ao alto custo dos experimentos de microarranjos, há na prática uma quantidade relativamente reduzida de dados disponíveis – em geral algumas poucas dúzias de pontos em séries temporais ou condições experimentais independentes –, enquanto a quantidade de genes envolvidos é da ordem de milhares. Um problema adicional é que os dados de expressão são extremamente ruidosos; erros de quantização podem atingir níveis de 30 a 50% (VINGRON & HOJEISEL,

1999). Este cenário tem levado a uma necessidade crescente de ferramentas computacionais capazes de capturar correlações não-lineares e lidar com interações estocásticas, sendo também ao mesmo tempo robustas o suficiente para operarem satisfatoriamente sob escassez de dados e informação ruidosa.

Entre as técnicas de modelagem existentes, capazes de representar e realizar inferências automáticas de uma rede gênica causal, as redes bayesianas (*Bayesian Networks* – BN) (PEARL, 1988) são consideradas dentre as opções mais atraentes. As redes bayesianas são naturalmente probabilísticas, possuem robustez a ruído e são sensíveis a correlações não-lineares. Com efeito, não é por acaso que as redes bayesianas são a metodologia mais adotada para engenharia reversa[1] de interações genéticas causais na literatura de bioinformática.

Redes bayesianas são também suficientemente flexíveis para serem adaptadas a domínios estáticos e dinâmicos. Redes estáticas têm como objetivo descobrir interações gênicas responsáveis pelos estados de equilíbrio do sistema. São interessantes, por exemplo, para analisar como genes interagem para dar origem a um estado fenotípico estável, como no caso de tecidos normais e cancerosos. A abordagem estática é um método eficiente para mapear os atratores da rede. Redes bayesianas dinâmicas (FRIEDMAN, *et al*., 1998) usam dados de séries temporais e também incorporam circuitos de realimentação, sendo portanto capazes de prover uma modelagem probabilística da dinâmica do processo sendo analisado.

Outra particularidade das redes bayesianas é que elas podem ser discretas ou contínuas. Em redes discretas, os níveis de expressão, originalmente contínuos, devem ser discretizados antes da análise. As relações de dependência condicionais podem então ser calculadas com exatidão através das tabelas de probabilidade condicional de Markov (*Conditional Probability Tables* – CPTs). A abordagem contínua, por sua vez, não envolve discretização. As relações condicionais são representadas por densidades marginais, calculadas com a ajuda de métodos aproximados de estimação de densidade de probabilidade.

[1] O processo de inferência de redes gênicas é mais conhecido na literatura de bioinformática como engenharia reversa, dado que a tarefa consiste em tentar compreender o funcionamento de um sistema já em operação através da manipulação desse sistema. O leitor habituado à nomenclatura "identificação de sistemas" deve atentar a essa particularidade.

Como os níveis de expressão gênica em uma célula pertencem naturalmente ao domínio contínuo, é de se esperar que as abordagens contínuas sejam mais indicadas para a reconstrução de redes reguladoras que as abordagens discretas. Genes operam grande parte do tempo em níveis de expressão intermediários e a discretização vai certamente levar à perda de informação relevante, que, aliás, já é bastante escassa. Surpreendentemente, contradizendo este raciocínio, a grande maioria dos estudos envolvendo inferência de redes gênicas utilizando redes bayesianas, sejam elas dinâmicas ou estáticas, fazem uso de variáveis discretizadas em vez de contínuas (FRIEDMAN *et al*., 1999; KHAN *et al*., 2002; ONG *et al*., 2002; PE'ER *et al*., 2001; PEÑA, 2004; SMITH *et al*., 2003; SPIRTES *et al*., 2000; YU *et al*., 2004; ZOU & CONZEN, 2005). De fato, redes bayesianas discretas são menos custosas computacionalmente quando o número de níveis discretos é pequeno, e são mais facilmente compreensíveis e implementáveis. No entanto, nenhum desses benefícios é suficiente para sustentar a escolha por discretização quando a capacidade de inferência é limitada pela quantidade reduzida de informação disponível; o que é definitivamente o caso para dados de expressão.

Provavelmente, a principal razão para evitar domínios contínuos está relacionada à necessidade de controlar a grande flexibilidade de algoritmos semi-paramétricos e não-lineares de estimação de densidade. Os métodos mais utilizados de estimação de densidade, como *Parzen windows*, *K-nearest neighbors* e *Gaussian kernels* (SCOTT, 1992), variam consideravelmente em performance sob pequenas modificações em seus parâmetros de regularização. Como já discutido em FRIEDMAN *et al*., (1999), a configuração desses parâmetros não é uma tarefa simples, embora propostas de redes bayesianas contínuas baseadas nesses tipos de métodos existam (HOFMANN & TRESP, 1996).

Mais recentemente, uma nova rede bayesiana contínua, baseada em mistura de modelos gaussianos e no algoritmo de maximização da esperança (*Expectation Maximization* – EM) (BILMES, 1998; BISHOP, 1995), foi proposta (DAVIES & MOOR, 2000). O algoritmo EM é uma abordagem muito eficaz. Quando usado em conjunto com algum critério de seleção de modelos, como BIC (*Bayesian Information Criterion*) ou AIC (*Akaike's Information Criterion*), ele se torna completamente automático em termos de ajuste paramétrico. De fato, EM tem sido extensivamente utilizado em aplicações recentes de bioinformática (como em PAN *et al*. (2003)), incluindo a reconstrução de redes

reguladoras (PERRIN *et al*., 2003). Um problema da estratégia EM é que os seus resultados dependem fortemente da inicialização, que é originalmente realizada de forma aleatória, e ela tende a produzir resultados diferentes a cada nova execução.

Além das limitações dos métodos de estimação de densidade descritos acima, existe uma outra – e certamente mais decisiva – dificuldade que surge devido à quantidade reduzida de dados disponíveis. De acordo com a teoria de regularização[2] (GIROSI *et al*., 1995), inferências não-lineares baseadas apenas em uma pequena quantidade de informação tenderão a sofrer de uma capacidade de generalização reduzida. Explicitamente, como apenas uma pequena quantidade de dados está disponível, métodos de regressão se tornarão tendenciosos em torno de pontos conhecidos, enquanto a predição em regiões mais desconhecidas se torna prejudicada. Este cenário é ainda mais agravado para o caso de expressão gênica, porque como os níveis de ruído são em geral muito elevados, até os pontos conhecidos tornam-se pouco confiáveis. Estratégias como métodos de *kernel*, que posicionam uma função de base radial sobre cada amostra disponível, irão certamente se sobre-ajustar aos dados. O mesmo é esperado para algoritmos como o EM, dada a sua grande flexibilidade.

Considerando esses aspectos desafiadores relativos aos algoritmos contínuos, é proposto aqui um novo método de estimação de densidade para redes bayesianas aplicado à reconstrução de redes gênicas no domínio contínuo. O método é particularmente projetado para lidar com conjuntos de dados pequenos, dando prioridade à generalização quando pouca informação está disponível. Essa proposta utiliza um algoritmo de sistemas imunológicos artificiais chamado ARIA (*Adaptive Radius Immune Algorithm*) (BEZERRA *et al*., 2005), que realiza uma compressão da informação, posicionando um número reduzido de protótipos (funções gaussianas) de acordo com a densidade de amostras no espaço. ARIA implementa um mecanismo adaptativo que é capaz de capturar a informação de densidade local e filtrar parte do ruído. Em uma segunda fase,

[2] Regularização é um conceito em estatística que está relacionado à suavidade de uma curva. No caso da estimação de densidade, funções de densidade de probabilidade mais regularizadas são funções de conformação mais suave. A regularização também está associada à capacidade de generalização de uma curva, no sentido de que curvas mais suaves atuam mais eficientemente na interpolação dos dados, sendo, portanto, em geral mais mais capazes de expressar o comportamento desejado em regiões onde há pouca informação disponível.

aprendizado supervisionado é utilizado para determinar automaticamente a variância das gaussianas, baseado no critério de máxima verossimilhança.

O método de estimação de densidade proposto, baseado no ARIA, será primeiramente comparado com o algoritmo EM quando ambos são aplicados em problemas de estimação de densidade com poucos dados. O propósito é avaliar como esses métodos se comportam em termos de desempenho sob circunstâncias forçosamente severas, mostrando que o ARIA é realmente capaz de evitar sobre-ajuste. Em um segundo experimento, uma rede bayesiana contínua, que utiliza o ARIA para estimação de densidade, é proposta e aplicada a dados de expressão artificiais, gerados por modelos sintéticos realistas de redes gênicas. Sua performance será comparada com a técnica mais utilizada para esse fim, as redes bayesianas discretas. Pretende-se mostrar como a discretização afeta a performance da inferência quando uma pequena quantidade de informação está disponível.

Embora análises experimentais com dados reais sejam desejadas, o conhecimento científico sobre estruturas de redes gênicas biológicas ainda é muito limitado. Experimentos com dados reais acabariam por se tornar restritos a análises e conclusões subjetivas (FRIEDMAN *et al*., 1999; VAN BERLO *et al*., 2003). Como argumentado em RICE *et al*. (2004), dado o conhecimento completo da estrutura verdadeira da rede em questão e a possibilidade de controle preciso da quantidade e qualidade dos dados, as redes sintéticas são ainda a melhor maneira de realizar uma comparação objetiva entre os diferentes métodos. De fato, dados artificiais têm sido amplamente utilizados para validação de outras técnicas de inferência na comunidade de bioinformática (KHAN *et al*., 2002; RICE *et al*., 2004; SMITH *et al*., 2003; YU *et al*., 2004).

2.3. Estimação de densidade

Estimação de densidade é a tarefa de inferência de uma função de densidade de probabilidade (*Probability Density Function* – PDF) baseada apenas nos dados gerados por essa função. Obviamente, como o número de dados disponíveis é em geral limitado e o processo de geração de dados é naturalmente estocástico, na maioria dos casos é praticamente impossível recuperar exatamente a PDF verdadeira. Na prática, entretanto, obter uma boa aproximação é usualmente possível e aceitável.

Nesta seção, o método de estimação de densidade proposto é apresentado. Em seguida, uma breve descrição do algoritmo de maximização da esperança é fornecida. Os experimentos computacionais realizados, comparando as duas técnicas, são apresentados e discutidos. Nesses experimentos preliminares, a natureza dos dados não tem relação com o problema de recuperação de redes gênicas, no qual as amostras representam genes e as condições experimentais a dimensão dos dados.

A. ARIA (Adaptive Radius Immune Algorithm)

ARIA é um algoritmo de Sistemas Imunológicos Artificiais (*Artificial Immune Systems* – AIS) (DE CASTRO & TIMMIS, 2002) originalmente proposto para clusterização baseada em densidade (BEZERRA *et al*., 2005). Usando idéias inspiradas em mecanismos do sistema imunológico, como o princípio da seleção clonal e a supressão da rede imunológica, ele realiza compressão dos dados através da geração de protótipos (anticorpos – Ab) que competem para o reconhecimento dos dados (antígenos – Ag) em um processo auto-organizado.

O procedimento de treinamento não-supervisionado pode ser resumido em três fases principais, como descrito a seguir (para uma descrição mais completa do algoritmo o leitor deve se referir a BEZERRA *et al*. (2005)):

1) *Maturação de afinidade*: os antígenos (dados) são apresentados aos anticorpos, e aqueles anticorpos que aprensentarem uma maior capacidade de reconhecer os antígenos, segundo alguma métrica, têm associados a si um grau de afinidade maior. Nessa fase, anticorpos sofrem hipermutação de forma a possivelmente melhor reconhecer os antígenos (interações do tipo Ag-Ab).
2) *Expansão clonal*: aqueles anticorpos que são mais estimulados (isto é, apresentam maior grau de afinidade) são selecionados para serem clonados. A rede imunológica cresce.
3) *Supressão da rede*: a interação dos anticorpos é quantificada, e se um anticorpo reconhece outro anticorpo, um deles é selecionado para ser removido do conjunto de protótipos (interações do tipo Ab-Ab).

Inicialmente, um número arbitrário de anticorpos é gerado em posições aleatórias. Aqueles anticorpos com maior grau de afinidade aos antígenos (isto é, aqueles protótipos

que têm uma distância pequena em relação aos dados) sofrem mutação em direção aos antígenos reconhecidos, em uma taxa proporcional à distância Ab-Ag. A seguir, os anticorpos que reconhecem muitos antígenos são clonados, isto é, eles geram cópias deles mesmos, fazendo assim o número total de anticorpos aumentar.

Para cada anticorpo *i* é associado um raio de supressão particular R_i. Se a distância entre dois anticorpos é menor que o raio de supressão de um deles (isto é, se há reconhecimento entre eles) aquele de maior raio é eliminado da população. Dessa forma, os anticorpos redundantes são suprimidos e apenas aqueles mais adaptados prevalecem.

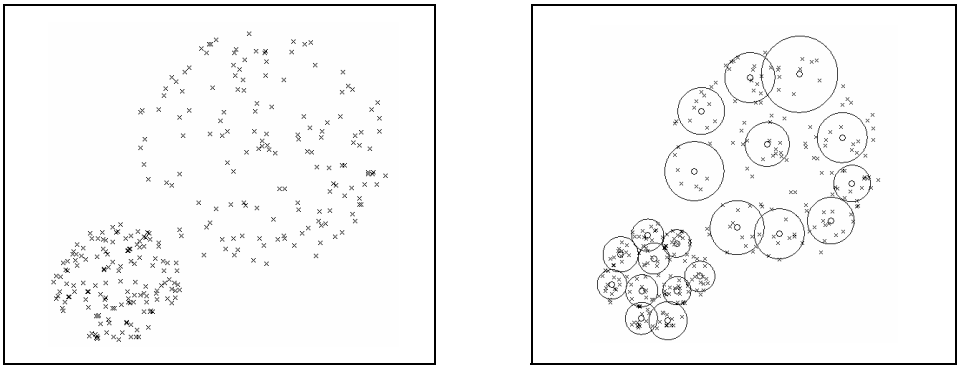
As três fases principais são repetidas seqüencialmente por muitas iterações, enquanto a taxa de mutação é gradualmente reduzida. Os agentes "imunológicos" devem interagir num processo auto-organizado que termina quando o número de anticorpos da população estabiliza e novas mutações não causam mais mudanças significativas no posicionamento dos anticorpos.

O recurso principal do algoritmo está relacionado ao raio de supressão adaptativo. Os valores independentes de R_i são escolhidos para serem inversamente proporcionais à densidade local em torno de cada anticorpo: em regiões densas, anticorpos terão raio pequeno, e em regiões esparsas, raios grandes. Dessa maneira, os anticorpos podem se aproximar mais uns dos outros em regiões de alta densidade, mas não podem ficar tão próximos entre si onde a densidade é baixa. Como resultado, no fim do processo a informação de densidade presente nos dados tende a ser maximamente preservada. Além disso, a distribuição de probabilidade dos anticorpos será semelhante à distribuição de probabilidade dos antígenos, mas como apenas a informação essencial é capturada, ruído e *outliers* acabam por serem filtrados.

Além disso, como uma conseqüência da auto-organização, o tamanho da população de anticorpos é também auto-regulado, isto é, ARIA automaticamente determina o número de protótipos que é necessário para uma representação de alta-qualidade. Todavia, é possível controlar o nível de especificidade da representação compactada através do ajuste do parâmetro *r*, que define o comprimento do menor raio de supressão da população (o raio do anticorpo localizado na região mais densa do espaço). Valores altos de *r* fornecem representações generalizadas e, indiretamente, menos anticorpos, enquanto valores pequenos definem representações mais acuradas e, conseqüentemente, mais anticorpos.

Essa destacada capacidade de auto-ajuste provê uma flexibilidade adicional ao procedimento de aprendizado do ARIA, que geralmente não é encontrado em algoritmos convencionais.

A Figura 2.1 fornece um exemplo ilustrativo da compressão de informação realizada pelo ARIA. Os dados na Figura 2.1(a) estão distribuídos em um espaço bidimensional e formam dois clusters com o mesmo número de pontos (200 amostras em cada cluster), mas de densidades diferentes. O ARIA foi executado para esse conjunto de dados e a configuração final dos anticorpos é mostrada na Figura 2.1(b), onde o centro de cada círculo representa a posição de um anticorpo e os raios dos círculos correspondem aos raios de supressão individuais.

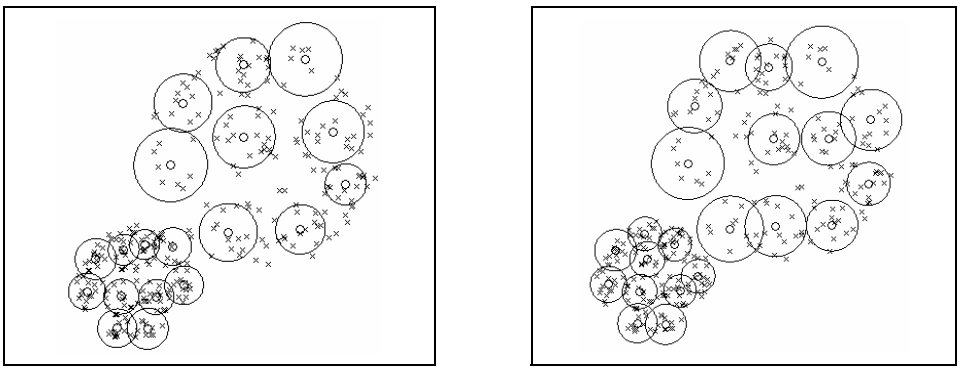


(a) **(b)**

Figura 2.1 (a) Conjunto de dados com dois clusters de densidades diferentes. (b) Posicionamento dos protótipos do ARIA.

Repare que o número de anticorpos em cada um dos clusters é o mesmo, e que eles estão muito mais próximos entre si na região do cluster de maior densidade do que na região do de menor densidade, mostrando que, embora os dados tenham sido compactados, a informação de densidade foi preservada dentro dos limites possíveis.

A Figura 2.2 mostra outras duas instâncias da execução do ARIA para o mesmo valor de *r*. Veja que embora a inicialização do algoritmo seja aleatória, a representação obtida não varia muito em cada caso.



(a) **(b)**

Figura 2.2 Duas novas execuções do ARIA para o problema dos dois clusters de densidades diferentes.

B. ARIA para estimação de densidade

Para aproximar a PDF verdadeira, a maioria dos métodos de estimação de densidade empregam misturas de gaussianas (ou modelos de mistura[3]) que, quando somadas, modelam uma função de probabilidade complexa. Misturas de gaussianas são de fato capazes de aproximar qualquer PDF (dado que o número de componentes é suficiente), mas entre muitas outras propriedades interessantes das funções gaussianas, a principal razão para elas serem as mais escolhidas está em sua tratabilidade analítica.

Aqui, a fase não-supervisionada é utilizada para definir os centros das funções gaussianas. Como as gaussianas são somadas para compor a PDF final, é imediato concluir que regiões densas do espaço vão precisar de mais gaussianas que regiões esparsas, o que propriamente coincide com o principal objetivo do ARIA. Isso não é necessariamente verdade para o caso em que as gaussianas podem ter pesos diferentes, pois uma gaussiana de peso elevado pode substituir muitas gaussianas de peso reduzido.

Para privilegiar generalização ao invés de especificidade, as gaussianas devem apresentar variâncias iguais, reduzindo assim o número de parâmetros do modelo. Além disso, seus pesos são também ajustados para serem iguais (precisamente $1/M$, onde M é o número de gaussianas), pois como é esperado que a densidade seja preservada, porções do

[3] Modelos de mistura são uma combinação de funções gaussianas, ou de uma outra função de probabilidade simples, que quando somadas modelam uma função de densidade de probabilidade complexa. Em estimação de densidade o emprego de um modelo de mistura está relacionado à determinação do número de gaussianas utilizadas, da altura (ou peso) de cada gaussiana, das suas aberturas (ou variâncias) e da posição no espaco dos seus centros de distribuição.

espaço de alta densidade serão naturalmente modeladas por um número maior de gaussianas, ao invés de poucas gaussianas com pesos elevados.

O papel dessas restrições é limitar a flexibilidade do modelo em situações em que os dados são escassos. Limitar a flexibilidade é uma maneira de obter PDFs mais regularizadas.

Para ajustar a variância das gaussianas, será adotada a equação 2.2, a ser apresentada na próxima subseção, a qual determina a variância de cada gaussiana de forma a maximizar a verossimilhança dos dados conhecidos em relação ao modelo. Como, no modelo proposto, as gaussianas devem apresentar variâncias iguais, o valor escolhido para esse parâmetro será a média de todas as variâncias individuais.

Mesmo para um modelo com tantas restrições de flexibilidade, a fase supervisionada relativa à determinação da abertura das gaussianas é ainda gulosa o suficiente para gerar um sobre-ajuste aos dados quando o número de componentes é relativamente alto. Para contornar esse problema, foi projetada uma estratégia de perturbação que consiste em adicionar um ruído gaussiano de baixa intensidade nos dados originais. Após a fase não-supervisionada, as amostras são perturbadas de forma a alterar ligeiramente a distribuição dos dados. Uma análise empírica mostrou que um ruído de desvio padrão de 0,05 é suficiente para os casos estudados aqui, considerando-se os dados normalizados. A equação 2.2 é então aplicada sobre esses pontos de maneira a maximizar sua verossimilhança. Essa estratégia mostrou ser bastante eficiente em aumentar a regularização das PDFs obtidas.

O método de estimação de densidade proposto pode ser resumido da seguinte forma:

1) Determine os centros das gaussianas usando ARIA;

2) Ajuste os pesos das gaussianas para $1/M$, onde M é o número de gaussianas;

3) Perturbe os pontos usando um ruído gaussiano de baixa intensidade;

4) Encontre as variâncias das gaussianas usando a equação 2.2 e tirando a média sobre o número de componentes.

C. Maximização da Esperança em Modelos de Mistura

Através da derivada da fórmula da verossimilhança de um modelo de mistura gaussiano, junto com um formalismo bayesiano e alguma manipulação algébrica, é possível deduzir expressões analíticas que determinam os parâmetros ótimos de um modelo de mistura em termos de maximização de verossimilhança (BILMES, 1998). As equações 2.1, 2.2 e 2.3 representam as fórmulas para os parâmetros ótimos que definem o modelo, onde μ_j, σ_j e w_j são a média, o desvio padrão e o peso da gaussiana j, respectivamente, e $j = 1,..., M$. Ainda nas equações, x_n representa o n-ésimo ponto dos dados, onde $n = 1,\ldots,N$, e $P(j \mid x_n)$ é a probabilidade a *posteriori* da gaussiana j dado x_n. A letra d representa a dimensão do conjunto de dados.

$$\mu_j^{novo} = \frac{\sum_n P^{antigo}(j \mid x_n) x_n}{\sum_n P^{antigo}(j \mid x_n)} \tag{2.1}$$

$$\left(\sigma_j^{novo}\right)^2 = \frac{1}{d} \frac{\sum_n P^{antigo}(j \mid x_n) \| x_n - \mu_j^{novo} \|^2}{\sum_n P^{antigo}(j \mid x_n)} \tag{2.2}$$

$$w_j^{novo} = \frac{1}{N} \sum_n P^{novo}(j \mid x_n) \tag{2.3}$$

Nas equações acima, as notações *novo* e *antigo* servem para denotar o procedimento iterativo a ser descrito mais adiante.

Para realizar o cálculo da probabilidade a *posteriori* $P(j \mid x_n)$, utiliza-se o teorema de Bayes da seguinte forma:

$$P(j \mid x) = \frac{p(x \mid j) \cdot P(j)}{p(x)} \tag{2.4}$$

Onde a distribuição da mistura, $p(x)$, é dada por:

$$p(x) = \sum_{j=1}^{M} p(x \mid j) \cdot P(j) \tag{2.5}$$

Essas equações não-lineares são de difícil otimização e não fornecem um método direto para o cálculo dos parâmetros. Entretanto, é possível contornar essa dificuldade aplicando-se um esquema iterativo que converge para um mínimo local, chamado maximização da esperança (*EM*, do inglês *Expectation Maximization*).

O algoritmo começa com parâmetros iniciais aleatórios, que chamaremos de "antigos". Em seguida, essa estimativa inicial é utilizada para calcular os "novos" valores dos parâmetros, para os quais o valor da função de verossimilhança deve aumentar. Após calcular todos os parâmetros, os valores "novos" se tornam agora "antigos", e o processo se inicia novamente. Esse esquema iterativo é repetido até que o algoritmo convirja. EM promove um método simples e prático de estimação dos parâmetros da mistura que evita as complexidades de algoritmos de otimização não-lineares.

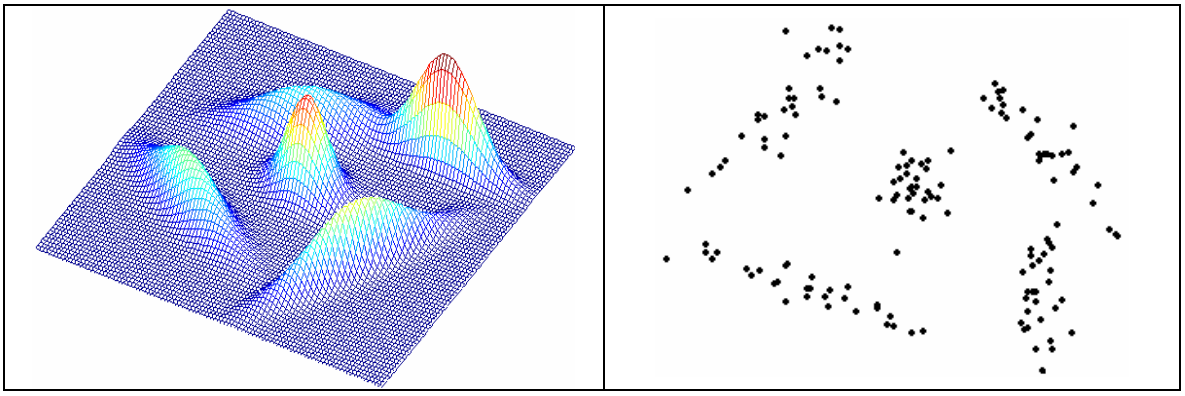
D. Experimentos em Estimação de Densidade Comparando ARIA e EM

Quando poucas amostras estão disponíveis, a estimação da verdadeira PDF se torna um problema muito difícil. Sob carência quantitativa de informação, é necessário dar prioridade à regularização das curvas obtidas; caso contrário, generalização e predição serão prejudicadas. É necessário abdicar da especificidade, ou as funções estimadas serão tendenciosas em torno dos pontos conhecidos.

Será analisada aqui a performance dos algoritmos ARIA e EM em dois problemas diferentes de estimação de densidade. A eficiência das técnicas é medida pela sua capacidade de generalização e predição, isto é, pela sua competência em maximizar a verossimilhança para pontos desconhecidos a priori.

No primeiro e mais simples problema analisado, a função de densidade de probabilidade original é composta de cinco gaussianas elípticas com matrizes de covariância distintas, fixadas em um espaço bi-dimensional. A Figura 2.3(a) mostra sua conformação tri-dimensional. Essa PDF foi utilizada para gerar 150 amostras, mostradas na Figura 2.3(b). Uma parcela de 80% dessas amostras foi escolhida aleatoriamente para

compor o conjunto de treinamento, e os 20% restantes foram utilizados para a fase de teste dos métodos.



(a) **(b)**

Figura 2.3 (a) Conformação tri-dimensional da PDF composta por gaussianas elípticas. (b) 150 amostras geradas pela PDF.

Para fazer uma comparação mais equilibrada entre os métodos, ARIA foi executado inicialmente para diferentes valores do parâmetro *r* e o número de gaussianas obtido para cada *r* foi então utilizado para inicializar o EM. Os valores de *r* usados variam de 0,003 a 0,06. Cada método foi executado 50 vezes e a média dos resultados é mostrada na Figura 2.4, onde o erro de verossimilhança (isto é, o negativo do logaritmo da verossimilhança) para os dados de treinamento e de teste por número de gaussianas são apresentados, junto com os seus respectivos desvios padrão da média.

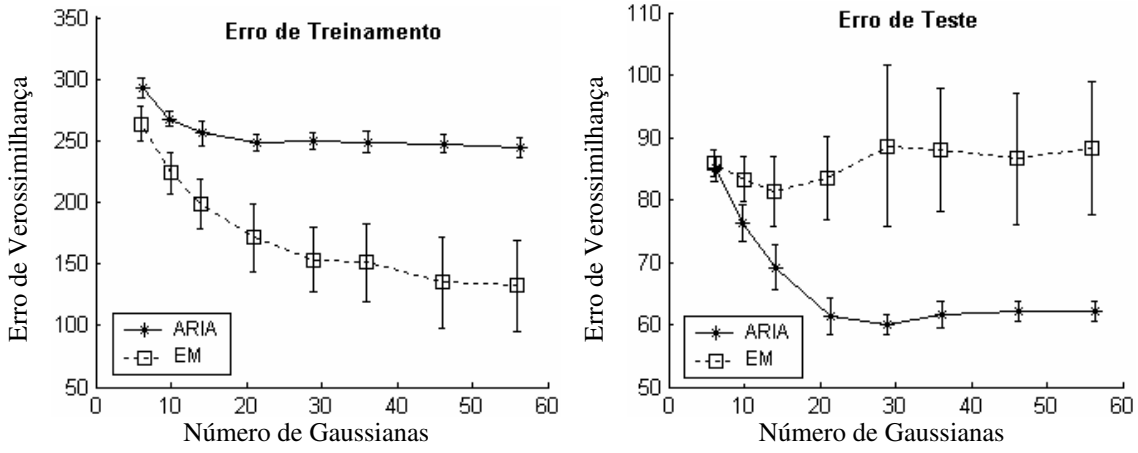


Figura 2.4 Desempenho do ARIA e do EM para o problema das gaussianas elípticas.

O erro de verossimilhança é dado pela equação 2.6:

$$E = -\ln L = -\sum_{n=1}^{N} \ln p(x_n) = -\sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{M} p(x_n \mid j) \cdot P(j) \right\}, \qquad (2.6)$$

onde E representa o erro, L representa a verossimilhança e as outras variáveis são as mesmas definidas na Seção 2.3.C.

Note na Figura 2.4 que o EM obteve um erro de treinamento menor que o do ARIA, mas na fase de predição a performance do ARIA foi muito superior. Esse resultado sugere uma evidência bastante forte de sobre-ajuste. Sendo mais flexível, o modelo gerado pelo EM se ajusta muito bem aos dados de entrada, reduzindo assim o erro de treinamento. No entanto, esse ajuste se torna específico demais a ponto de o modelo perder a capacidade de generalização, resultando num elevado erro de teste quando sujeito a dados desconhecidos. Veja também que, à medida que a complexidade do modelo cresce (o número de gaussianas aumenta), o erro de treinamento tende a reduzir, pois a flexibilidade do modelo de misturas também aumenta. Como conseqüência, há uma perda de regularização das curvas, e uma predição de baixa qualidade é obtida.

Por outro lado, as restrições de flexibilidade impostas ao modelo gerado pelo ARIA reduzem o sobre-ajuste aos dados, aumentando assim a sua capacidade de generalização (para o problema em questão) e levando a um erro de teste menor. Note na curva de teste que o número de gaussianas é inicialmente insuficiente, e que o erro de predição é gradualmente reduzido até que o número de gaussianas seja o bastante para uma representação de boa qualidade. À medida que a flexibilidade aumenta ainda mais, a curva de erro permanece estável, pois o mecanismo de perturbação força a abertura das gaussianas a ser relativamente alta, evitando assim o sobre-ajuste mesmo sob essas condições. Isso pode ser observado também na curva de erro de treinamento.

Outro ponto interessante nos gráficos é que os desvios padrão no erro obtido pelo algoritmo EM são muito maiores que aqueles obtidos pelo modelo gerado pelo ARIA. Isso certamente ocorre porque a performance do EM depende muito de sua inicialização, enquanto o ARIA, que também é inicializado aleatoriamente, é capaz de encontrar representações aproximadamente equivalentes a cada nova execução.

O segundo problema analisado nesta seção consiste do conjunto de dados Iris. Esse conjunto de dados é amplamente utilizado na comunidade de aprendizado de máquinas para validação de técnicas de clusterização e classificação. Os dados são compostos de 150 pontos e quatro atributos, representando três espécies de plantas. Desta vez o problema de estimação se torna mais difícil, porque de acordo com o princípio da "maldição da dimensionalidade" (BELLMAN, 1961), à medida que o número de dimensões cresce linearmente, a quantidade de dados deve crescer exponencialmente de modo a manter a mesma densidade amostral. Como a dimensão dos dados dobrou de valor e o número de pontos foi mantido o mesmo do problema anterior, é possível concluir que a informação disponível foi drasticamente reduzida.

O procedimento experimental empregado é o mesmo: 80% de dados para treinamento, 20% para teste e cada algoritmo foi executado 50 vezes (os dados de treinamento e teste são redefinidos a cada execução). Os valores de r utilizados variam de 0,02 a 0,06. Os resultados são apresentados na Figura 2.5.

As curvas obtidas são qualitativamente similares às da Figura 2.4. Novamente, o modelo gerado pelo ARIA atingiu um desempenho muito superior ao EM em termos de erro de teste, embora a distância relativa entre as curvas tenha sido reduzida para este experimento.

Veja que o erro de teste do ARIA decresce mais suavemente, desta vez, à medida que a complexidade aumenta, e atinge seu mínimo quando o número de gaussianas é em torno de 30. Para este problema, os desvios padrão obtidos pelo EM parecem ter aumentado. De fato, a magnitude dos desvios observados nos gráficos são certamente inaceitáveis para a maioria das aplicações.

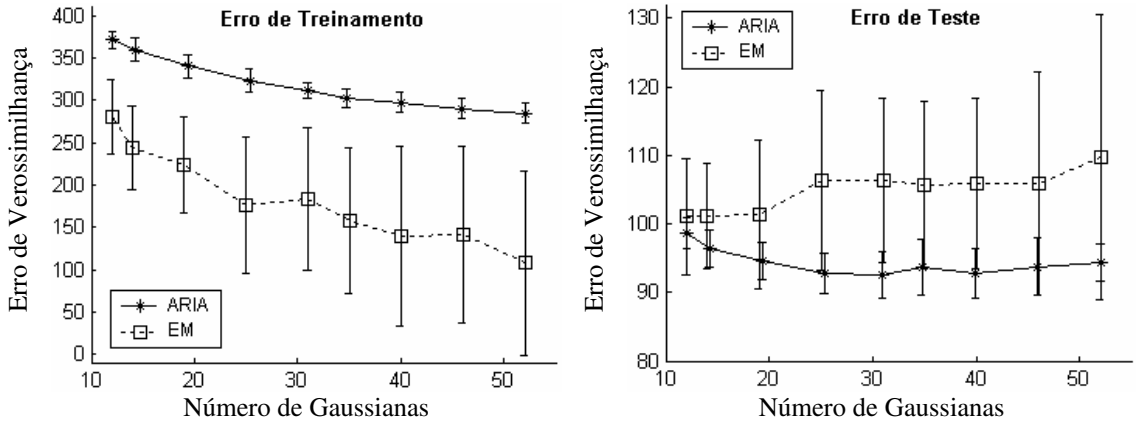


Figura 2.5 Desempenho do ARIA e do EM para o conjunto de dados Iris.

2.4. Recuperação Redes Gênicas

A. Modelagem com Redes Bayesianas

O processo de identificação de sistemas com redes bayesianas ocorre em dois níveis distintos. O primeiro, e mais baixo nível, consiste em determinar as correlações locais conjuntas entre as variáveis utilizando métodos de estimação de densidade ou de regressão. O segundo, e mais alto nível, corresponde à busca por uma estrutura de rede que descreve a maneira pela qual as variáveis mais correlacionadas interagem entre si em termos de causalidade.

A rede bayesiana é dividida em módulos chamados famílias, cada uma delas consiste de um nó (variável) e os pais daquele nó, isto é, aquelas variáveis que afetam o nó filho diretamente. Considere então um grafo acíclico direcionado *G* definindo a estrutura de uma rede bayesiana. A verossimilhança de uma família com variável filha X_i e um conjunto de pais $\vec{\Pi}_i$ é dada pela probabilidade marginal $P_i\left(X_i \mid \vec{\Pi}_i\right)$. Como a densidade de probabilidade marginal não é disponível diretamente, é necessário primeiramente estimar a probabilidade conjunta $P_i\left(X_i, \vec{\Pi}_i\right)$ usando um algoritmo de estimação de densidade. A seguir, a densidade marginal $P_i\left(\vec{\Pi}_i\right)$ é calculada baseado na probabilidade conjunta, que no caso de uma mistura de gaussianas pode ser obtido com certa facilidade: o número de componentes da forma de funções gaussianas dessa densidade marginal e também os seus

pesos serão os mesmos do modelo conjunto. Suas médias e matrizes de covariância também serão as mesmas, mas todos os elementos correspondentes à variável X_i são removidos.

A verossimilhança dos dados para um módulo é então dada por $P_i\left(X_i \mid \vec{\Pi}_i\right) = P_i\left(X_i, \vec{\Pi}_i\right) / P_i\left(\vec{\Pi}_i\right)$, e a verossimilhança da rede se torna:

$$P\left(\vec{X}\right) = \prod_{i=1}^{N} P_i\left(X_i \mid \vec{\Pi}_i\right) = \prod_{i=1}^{N} \left(\frac{P_i\left(X_i, \vec{\Pi}_i\right)}{P_i\left(\vec{\Pi}_i\right)} \right), \tag{2.4}$$

onde *N* é o número de variáveis.

Baseado nessa fórmula, uma heurística de busca é então empregada para encontrar a estrutura de rede que maximiza o valor da verossimilhança. Vários algoritmos de aprendizado de estrutura existem, como o *greedy hill climbing*, *beam search*, *simulated annealing* e o *best first search* (veja VAN BERLO *et al*. (2003) para uma revisão dessas técnicas). A maioria deles consiste basicamente em iniciar com uma rede aleatória (ou uma população delas) e aplicar operadores de mutação para aumentar iterativamente sua verossimilhança.

Guiar a busca baseando-se apenas no critério de máxima verossimilhança (ML), no entanto, é geralmente uma opção arriscada. O ML vai tender sempre a favor dos modelos mais complexos e mais especificamente ajustados aos dados conhecidos, levando provavelmente a resultados tendenciosos. Mais uma vez, levar em conta a generalização é uma escolha mais razoável, e nesse caso evitar complexidade (e especificidade) está relacionado a restringir o número de conexões da rede.

Como alternativa a utilizar exclusivamente a informação de verossimilhança, uma estratégia mais interessante é fazer uso de um critério de seleção de modelos, como *Bayesian Information Criterion* (BIC) (SCHWARZ, 1978), ou *Akaike's Information Criterion* (AIC) (AKAIKE, 1974), que penalizam a complexidade. As fórmulas do BIC e AIC são mostradas abaixo nas equações 2.7 e 2.8, respectivamente. O primeiro termo do lado direito das duas equações é exatamente o logaritmo da função de verossimilhança para um conjunto de parâmetros θ igual a $\hat{\theta}$, normalizado pelo número de pontos *n*. O segundo termo consiste de um coeficiente de penalização de complexidade, onde *k* é o número de parâmetros do modelo. Note que, à medida que o número de parâmetros cresce, isto é, a

complexidade aumenta, o coeficiente de penalização também cresce. Veja também que o BIC implementa uma penalidade para complexidade maior que o AIC.

$$BIC = \frac{\ln L\left(x;\theta = \hat{\theta}\right)}{n} - \left[\frac{\ln(n)}{2}\right] * \frac{k}{n} \tag{2.7}$$

$$AIC = \frac{\ln L\left(x;\theta = \hat{\theta}\right)}{n} - \frac{k}{n} \tag{2.8}$$

Os valores BIC e AIC serão, portanto, responsáveis por limitar o número de parâmetros da rede, os quais crescem com o número de conexões.

B. Número de Amostras *versus* Número de Genes

Quando se trata de dados de expressão gênica e redes reguladoras, uma questão intrigante deve ser analisada: quantas amostras de dados são necessárias para inferir com confiabilidade uma rede de *n* genes? Em outras palavras, quantas amostras por gene são realmente necessárias? Dado que o número de genes em um experimento de mirroarranjos é geralmente tremendamente maior que o número de amostras disponíveis, este é certamente um aspecto relevante a ser considerado (DAVIES & MOORE, 2000).

Fornecer uma resposta precisa a essa pergunta é extremamente difícil, ou mesmo impossível. O número de pontos necessários vai depender de vários fatores, como os tipo de não-linearidade envolvidos, a qualidade dos dados em termos de nível de ruído, a representatividade dos experimentos (por exemplo, se os experimentos cobrem um conjunto significativo de condições experimentais consideravelmente diferentes), entre outros. Balancear todas essas questões simultaneamente está longe de ser uma tarefa simples.

Contudo, uma análise mais criteriosa permitiria observar que a questão acima é, na verdade, conceitualmente errônea. O número de genes sob consideração não é a verdadeira variável sendo limitada aqui. Uma pergunta mais adequada seria: quantas conexões por gene podem ser inferidas com confiabilidade, dado um número fixo de amostras? Mais especificamente, para o caso das redes bayesianas, quantos pais uma família da rede pode ter, dado um número limitado de pontos?

Quando se tenta recuperar a rede original, os dados são efetivamente usados para determinar as correlações entre as variáveis, que são dadas pela verossimilhança das famílias da rede. Calcular essa verossimilhança implica em estimar a densidade conjunta das variáveis de uma família. Se a família tem um pai, a PDF (do inglês *Probability Density Function*) conjunta estimada terá duas dimensões. Se tiver dois pais, a PDF conjunta terá três dimensões. Se tiver três pais, quatro dimensões, e assim sucessivamente. Entretanto, à medida que o número de dimensões cresce, a quantidade de informação disponível para estimar a densidade conjunta decresce exponencialmente (como discutido anteriormente na Seção 2.3.D). Isso significa que a principal questão está relacionada a determinar até que ponto as densidades conjuntas multidimensionais podem ser estimadas dado um número fixo de amostras. Uma vez que isso foi definido, o tamanho da rede relativo ao número de variáveis pode, em teoria, ser qualquer. A partir desse ponto, a quantidade de dados não influencia mais, embora a demanda de esforços para os algoritmos de busca irá aumentar com o aumento do número de genes, já que a taxa de crescimento do espaço de busca é mais rápida que um crescimento exponencial.

Portanto, métodos de estimação de densidade têm um papel fundamental em tentar maximizar a utilização dos dados. Quanto mais a informação disponível é propriamente utilizada, maior o grau de conectividade aceitável e mais confiáveis serão as redes inferidas.

C. Redes Reguladoras Sintéticas

Agora vamos abordar a questão de definir redes sintéticas realistas, as quais serão utilizadas nos experimentos computacionais de inferência de redes. O principal ponto a ser considerado aqui é tentar simular redes *in silico* com níveis de complexidade (em termos de quantidade e qualidade) que podem ser encontrados em redes reais. Nessas condições, é possível esperar que uma técnica de modelagem que desempenha bem no primeiro caso, também terá um bom desempenho no segundo.

A abordagem sintética empregada aqui se concentra em cinco pontos principais:

1) *Conectividade*: redes gênicas são esparsas, com genes sendo regulados por um número limitado de outros genes, de acordo com a distribuição da lei da potência (JEONG *et al*., 2001);

2) *Não-linearidade*: interações reguladoras são essencialmente não lineares, seguindo funções que saturam quando o gene é sub ou sobre-expressado;
3) *Funcionalidade lógica*: os mecanismos de controle gênicos podem ser cumulativos (OR) ou multiplicativos (AND);
4) *Ruído e estocasticidade*: dados de expressão são extremamente ruidosos e as relações genéticas são naturalmente estocásticas;
5) *Dados escassos*: o número de dados disponíveis é muito limitado, geralmente no máximo 50 a 100 experimentos.

Para garantir redes gênicas esparsas, a conectividade das redes foi limitada de forma que cada nó da rede tenha no máximo dois pais, o que é consistente com o pequeno número de amostras disponíveis. Infelizmente, esta restrição implica que a lei da potência não pode ser estritamente seguida para redes com muitos genes.

As interações gênicas são modeladas pela função sigmoidal descrita na equação 2.9, assim como empregado em (WEAVER *et al*., 1999), onde x_i é o nível de expressão do gene *i*, r_i é o estado regulador, e α_i e β_i são duas constantes específicas para cada gene, que definem a inclinação da curva e sua média, respectivamente:

$$x_i = \frac{1}{1+\exp(-\alpha_i r_i - \beta_i)} \tag{2.9}$$

O estado regulador do gene *i* é dado por:

$$r_i = \sum_j w_{i,j} u_j \quad \text{ou} \quad r_i = \prod_j w_{i,j} u_j \ , \tag{2.10}$$

onde u_j é o nível de expressão do gene *j* que causa *i* e $w_{i,j}$ define a força da interação. Dessa forma, diferentemente de outras abordagens que consideram somente relações reguladoras aditivas, foram levadas em conta funcionalidades do tipo OR e AND.

Dados de expressão são intrinsecamente muito ruidosos devido às condições experimentais para obtenção de dados de microarranjos e também devido a flutuações estocásticas nos processos celulares. Para se ater a um cenário mais realista, foi introduzida uma razão sinal-ruído (*Signal to Noise Ratio* – SNR) de 50%, isto é, o desvio padrão do

erro gaussiano empregado é metade do desvio padrão do sinal. Não é do conhecimento do autor outra abordagem sintética que utilize um nível de erro tão alto.

Os dados de expressão das variáveis independentes (nós da rede que não possuem pais) são gerados por distribuições normais com desvios padrão aleatórios e centros também determinados aleatoriamente. Com efeito, a natureza dessas distribuições não importa realmente se for garantido que os dados cobrem com representatividade um intervalo de valores significativo.

A topologia da rede é gerada aleatoriamente, seguindo as restrições de conectividade, e 20% das variáveis são selecionadas para serem independentes. O tipo lógico das interações é determinado aleatoriamente para cada família e os coeficientes das funções sigmoidais, α e β, são definidos como inteiros aleatórios no intervalo [–10,+10]. Após gerar os dados das variáveis independentes, os valores obtidos são utilizados para determinar os dados para as variáveis dependentes, seguindo as conexões da rede.

D. Experimentos

Os experimentos foram realizados em duas redes simuladas: uma rede menor, com apenas 6 genes, com a qual é possível fazer uma análise mais detalhada, e uma maior, com 20 genes. A rede bayesiana contínua utilizando o ARIA para estimação de densidade foi comparada com uma rede bayesiana discreta em sua capacidade de recuperar a estrutura da rede original baseado somente nos dados gerados por ela. A heurística de busca utilizada aqui para o aprendizado de estrutura (a qual será a mesma para os dois modelos de redes bayesianas) será o *greedy hill climbing* – uma busca gulosa reiterada baseada no máximo gradiente –, pois essa técnica apresentou os melhores resultados quando comparada com outros algoritmos em VAN BERLO *et al*. (2003). O funcionamento do *hill climbing* é semelhante ao do algoritmo K2, apresentado no Apêndice, com a exceção de que aqui inicia-se com uma rede aleatória, em vez de uma rede totalmente sem arcos, e se admite remoção de arcos também.

Para o processo de discretização das variáveis, foi empregado o método proposto em PEÑA (2004), que utiliza o algoritmo de clusterização *k-means* para determinar três níveis de expressão. Trata-se de um método avançado de discretização, e também foi utilizado em VAN BERLO *et al*. (2003).

Para a rede contínua, o critério de seleção de modelos AIC foi empregado, pois como é esperado que o ARIA já tenha tratado em parte o problema de sobre-ajuste gerado pelo critério de máxima verossimilhança, não é necessário dar uma penalidade muito forte para a complexidade, como ocorre no caso do critério BIC. Para o modelo discreto, o critério BIC de seleção de modelos foi aplicado. Testes preliminares mostraram que o AIC deixa esse modelo excessivamente flexível, resultando em redes com muitas conexões incorretas.

A primeira estrutura de rede sintética, com 7 conexões, é mostrada na Figura 2.6, onde os tipos de funcionalidade conjunta também são destacados.

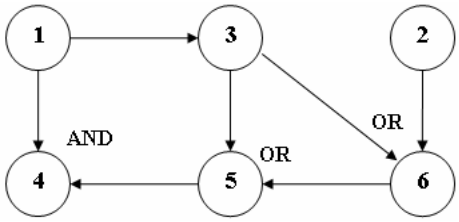


Figura 2.6 Estrutura da primeira rede sintética, com 6 genes.

O *hill climbing* foi executado 50 vezes para cada modelo usando quantidades diferentes de dados. Alguns testes empíricos semelhantes àqueles apresentados na Seção 2.3.D, realizados utilizando algumas variáveis da rede original selecionadas aleatoriamente, indicaram um valor de r de 0,007 para o ARIA quando o número de amostras é 50 e $r = 0{,}01$ para mais de 50 amostras. Os resultados, apresentados na Tabela 2.1 a seguir, incluem o número de conexões corretas, incorretas e o total de conexões encontradas para a melhor rede de cada teste.

Fica claro pela Tabela 2.1 que a abordagem contínua obteve uma performance muito superior àquela da abordagem discreta. Com ela, 70 amostras são suficientes para recuperar a estrutura verdadeira da rede, enquanto a rede bayesiana discreta necessitou de 2000 amostras. Além do mais, com apenas 50 amostras o modelo contínuo encontra 5 arcos da rede original, quase a rede inteira, enquanto a discreta encontra apenas um.

Tabela 2.1 Resultados para a rede de 6 genes.

Nº de Amostras	Contínua			Discreta		
	incorretas	**corretas**	**total**	**incorretas**	**corretas**	**total**
50	0	5	5	0	1	1
70	0	7	7	0	3	3
100	0	7	7	0	2	2
200	0	7	7	0	4	4
1000	0	7	7	0	6	6
2000	0	7	7	0	7	7

Com mais de 200 amostras, a rede discreta identifica apenas interações simples, necessitando de 1000 amostras para que relações conjuntas sejam detectadas. Esse resultado peculiar parece ser causado pela penalidade de complexidade imposta pelo BIC, a qual dá preferência para interações simples. Pode também parecer inconsistente que para 70 amostras a rede discreta encontra três arcos da rede, enquanto para 100 amostras, apenas 2 arcos são encontrados. Entretanto, isso acontece porque em cada teste instâncias diferentes dos dados são utilizadas.

O próximo experimento computacional foi realizado com uma rede de 20 genes e um total de 31 conexões. O parâmetro *r* foi configurado para 0,01 e o algoritmo *hill climbing* foi executado 1000 vezes para cada método. Um total de 100 amostras foi utilizado.

O modelo discreto encontrou apenas 5 conexões, todas elas arcos da rede verdadeira, o que é consistente com os resultados obtidos no experimento anterior. A rede bayesiana contínua, entretanto, detectou 29 arcos: 23 pertencem realmente à rede original, mas 6 são conexões identificadas incorretamente.

Embora o modelo contínuo tenha encontrado 23 conexões corretas (o que é muito mais do que as 5 do modelo discreto), a identificação de conexões incorretas é geralmente indesejada pois pode revelar interações genéticas que não existem realmente. É mais interessante, nesse caso, produzir grafos que possuam um número reduzido de conexões, porém que apresentem maior consistência.

Vamos então analisar mais de perto os resultados obtidos de forma a tirar conclusões adicionais. Primeiramente, foi notado que nas 1000 execuções do *hill climbing* a rede bayesiana discreta convergiu para um número bastante reduzido de grafos de estrutura

diferente, enquanto a contínua convergiu para uma grande variedade de estruturas alternativas. Isso sugere que a superfície de possíveis valores do critério BIC discreto tem um número pequeno de ótimos locais, enquanto a superfície de valores do AIC contínuo produz um número maior de ótimos locais. Mais uma vez, o resultado particular do modelo discreto pode ter sido causado pela forte penalidade imposta pelo BIC, a qual limita o número de conexões admitidas, e apenas aquelas que correspondem a interações verdadeiramente fortes permanecem. De fato, quando a abordagem contínua do ARIA foi aplicada juntamente com o critério BIC, resultados semelhantes foram encontrados, mas com a detecção de 10 conexões corretas, ao invés de 5, e nenhuma incorreta.

Outro aspecto a ser destacado é que a melhor rede encontrada pela busca *hill climbing* tem um valor AIC menor que o da rede verdadeira. A primeira obteve um valor de −28,82, enquanto a última, −28,37. Isso significa que a rede verdadeira de fato existe como um (provável) ótimo global, mas o *hill climbing* não foi capaz de encontrá-lo. De fato, há ainda uma diferença significativa entre o valor AIC da melhor rede encontrada e aquele da rede verdadeira, considerando a pior rede encontrada pelo *hill climbing*, que obteve AIC igual a −30,41. A conclusão é que o modelo gerado pelo ARIA foi capaz de evidenciar corretamente as correlações mais fortes entre os genes, mas como o número de dados é muito pequeno, a diferença em valores AIC para variáveis correlacionadas e descorrelacionadas é relativamente pequena, gerando assim uma superfície de valores AIC difícil de ser otimizada globalmente. O *hill climbing* demonstrou ser inadequado para essa tarefa, pois ele é incapaz de evitar ótimos locais, e encontrar a melhor rede se baseia principalmente na chance de inicializar o algoritmo já em um ponto muito promissor do espaço de busca. Para redes reguladoras de 20 ou mais genes, em que o número de estruturas possíveis é extremamente alto, técnicas de otimização menos sensíveis à inicialização devem ser adotadas.

2.5. Discussão

Neste capítulo, um novo algoritmo de estimação de densidade para redes bayesianas aplicadas à inferência de redes gênicas foi proposto. O modelo trabalha em domínio contínuo e é capaz de lidar com informação insuficiente e níveis elevados de ruído, sendo, portanto, especialmente adequado para dados de expressão gênica. Experimentos realizados

com redes simuladas realistas mostraram que o método proposto é capaz de identificar corretamente variáveis correlacionadas com poucas amostras, enquanto o método discreto, o mais utilizado na literatura, necessita em torno de 30 vezes mais amostras para atingir os mesmos resultados.

Foi verificado também que a fase de aprendizado de estrutura tem uma importância enfatizada quando se lida com poucos dados. Isso acontece porque a diferença entre variáveis correlacionadas e descorrelacionadas, quando percebida, é relativamente pequena, levando a uma superfície de busca difícil de otimizar. Conseqüentemente, o algoritmo *hill climbing*, incapaz de evitar ótimos locais, teve uma performance ruim. Esses resultados contradizem o sentimento comum existente na literatura de que heurísticas de busca mais sofisticadas não são realmente necessárias. Baseado nos resultados obtidos aqui, se torna difícil não objetar simulações computacionais como as realizadas em VAN BERLO *et al*. (2003), onde apenas 5% de ruído é introduzido nos dados e as correlações entre as variáveis são fortíssimas (acima de 90%) e consideradas como originalmente discretas (isto é, não há perda de informação por discretização); um cenário, sob todas as considerações práticas, ideal. Nos experimentos realizados aqui, as correlações, ao contrário, tendem a ser muito fracas, dado o elevado nível de ruído empregado e as variações aleatórias na conformação das curvas de regulação consideradas.

Reforça-se que a análise de desempenho apresentada aqui não poderia ser imediatamente obtida através de dados biológicos reais, pois a estrutura de redes reais não é perfeitamente conhecida. Diferentemente de outros tipos de análise, como clusterização ou classificação, não há um conjunto de dados padrão amplamente utilizado na literatura para validação de novas técnicas de inferência de redes gênicas. Como resultado, cada nova abordagem propõe sua própria metodologia de validação, e não há um senso comum entre os especialistas sobre quais são as melhores dentre elas. Enquanto o conhecimento a respeito de redes reais permanecer muito limitado, procedimentos padrões de simulação precisam ser definidos. A sugestão proposta é que esses procedimentos sejam focados nas mesmas características de complexidade consideradas aqui, tais como não-linearidades, conectividade esparsa, altos níveis de ruído, estocasticidade e quantidade de dados reduzida.

Capítulo 3

Redes Gênicas Artificiais

Resumo – O projeto evolutivo de sistemas artificiais é uma tendência crescente no estudo do funcionamento das redes gênicas e protéicas. Embora consistam em abstrações matemáticas das redes reais, as redes artificiais promovem perspectivas de investigação inteiramente novas, dado que todos os aspectos do sistema podem ser manipulados e/ou armazenados para análises futuras. Considerando este cenário, neste capítulo um modelo conexionista das redes gênicas e protéicas é proposto e, a partir desse modelo, sistemas artificiais capazes de realizar tarefas dinâmicas complexas são projetados por meio de um procedimento evolutivo. No modelo proposto, a evolução ocorre através de mutações estruturais, nas quais reações bioquímicas aleatórias – representadas como estruturas em grafo direcionado com conexões funcionais – são adicionadas ao sistema, prevalecendo ou não de acordo com a pressão seletiva. O modelo conexionista é contrastado com abordagens já existentes na literatura e é avaliado em termos de sua capacidade de evoluir comportamento de quimiotaxia em bactérias artificiais móveis expostas a substâncias químicas em um ambiente virtual. As redes reguladoras obtidas são analisadas considerando a relação entre estrutura e dinâmica. Os resultados dos experimentos mostram que o modelo proposto é capaz de reproduzir características observadas em organismos reais simples, e a análise e manipulação das redes obtidas fornecem uma explicação para a emergência dessas características.

3.1. Considerações Iniciais

A modelagem computacional está se tornando uma metodologia fundamental no processo de investigação dos sistemas biológicos. Modelos em computador têm a vantagem de serem específicos (isto é, pode-se modelar apenas os aspectos de interesse do sistema), podem ser manipulados arbitrariamente com facilidade e, dependendo do tipo da

modelagem empregada e da magnitude do sistema, é possível obter respostas rápidas a experimentos que em laboratórios de biologia (*in vivo*) demorariam dias ou semanas.

Desta motivação têm resultado algumas propostas de modelagem de redes gênicas, como as redes booleanas e a modelagem com equações estocásticas, cada uma delas empregando um enfoque próprio. Entretanto, como será discutido mais adiante na Seção 3.3 deste capítulo, as abordagens existentes desconsideram em sua modelagem grande parte das características essenciais das redes gênicas como sistemas de processamento de informação, tornando assim a modelagem abstrata fundamentalmente incapaz de representar propriedades de interesse observadas em organismos vivos. Alguns desses aspectos fundamentais podem ser sumarizados como a seguir:

- Uma vez que as redes gênicas são sistemas de funcionamento integrado – assim como os sistemas vivos em geral –, é de se esperar que as interações com o ambiente sejam de fundamental importância para a sua constituição, organização e funcionamento. Considerar um sistema vivo em isolamento vai contra a concepção moderna de sistema vivo (SCHNEIDER & KAY, 1994), isto é, um sistema em não-equilíbrio, aberto à troca de matéria e informações com o ambiente.
- Pesquisas em teoria de redes têm mostrado que os sistemas vivos possuem uma estrutura em rede bastante complexa e organizada, e que essa estrutura tem importância fundamental na dinâmica desses sistemas. A conjectura é que não se pode entender o funcionamento desse tipo de sistema através da análise da estrutura apenas, ou da dinâmica apenas.
- Um outro aspecto relevante é a funcionalidade do sistema, representada pelo conjunto de tarefas executadas pelo sistema modelado. Só é possível avaliar o processamento de informação caso o sistema efetivamente processe informação e realize alguma operação em função disto. Se não há funcionalidade, então não há processamento útil de informação. É como um sistema mecânico que recebe energia mas dissipa tudo em calor, sem realizar trabalho algum.
- Uma última questão está relacionada à integração gene-proteína no processamento de informação. Ao contrário do que é comumente adotado, não existe separação entre rede gênica e rede protéica. Uma vez que as proteínas executam o papel de regular a ação gênica e que as proteínas interagem com outras proteínas

constantemente (ou seja, existem proteínas que "regulam" as proteínas reguladoras), não faz sentido considerar apenas o processo de síntese/regulação no controle da expressão gênica. As interações proteína-proteína têm papel fundamental no processamento de informação celular e não podem ser desvinculadas das interações gene-proteína como sendo um caso à parte. Por conseguinte, elas não devem em princípio ser ignoradas na modelagem computacional.

As abordagens existentes de redes gênicas artificiais ignoram a maioria das considerações feitas acima, o que pode ser considerado como uma das razões pelas quais nenhum desses modelos tem se mostrado satisfatório na explicação de como se dá o processamento de informação celular; ou, mais especificamente, de que maneira uma célula é capaz de interpretar e reagir apropriadamente a mensagens do ambiente através de seu sistema de regulação constituído de genes e proteínas.

Neste capítulo, é apresentada uma nova proposta para modelagem de redes gênicas e protéicas, que será chamada aqui de modelagem conexionista. Essa proposta de modelo, juntamente com os procedimentos de simulação adotados, compõe uma metodologia de modelagem e investigação que tenta incorporar simultaneamente todos os aspectos descritos acima, diferindo assim significativamente das abordagens propostas até agora.

A Seção 3.2, a seguir, apresenta uma introdução em que o leitor é conduzido por uma linha de raciocínio que motiva o trabalho apresentado, seguida por uma descrição do conteúdo do capítulo.

3.2. Motivação e Posicionamento da Proposta

As investigações sobre o funcionamento das redes gênicas e protéicas têm se concentrado em descrições detalhadas de mecanismos celulares e circuitos de regulação genética específicos. Embora muitas características globais dessas redes já tenham sido elucidadas – como a distribuição em lei da potência de seu grau de conectividade (JEONG *et al*., 2000) e sua estrutura hierárquica modular (RAVASZ *et al*., 2002) –, poucos avanços foram efetivamente alcançados em termos de uma perspectiva sistêmica. Até agora não está claro como essas características de grande escala estão relacionadas a vias metabólicas ou circuitos reguladores específicos, ou mesmo se essas análises detalhadas serão capazes de

fornecer contribuições significativas para a compreensão das propriedades emergentes de tais sistemas.

Não obstante, a metodologia reducionista aparece como uma das principais alternativas de pesquisa quando se trata da análise de organismos reais, muito embora ela apresente limitações evidentes. A escala das redes gênicas e de proteínas para os organismos mais simples na Terra é grande demais para uma investigação holística, e como esses sistemas funcionam como entidades integradas, partes menores não podem ser apropriadamente isoladas para estudo. Além disso, as estruturas básicas de funcionamento são as mesmas para todas as forma de vida (JACOB, 1998), e não há muita evidência de projetos alternativos, o que seria de fundamental importância como material para realizar análises comparativas.

Uma alternativa válida a este cenário é tentar criar formas de vida artificiais, que correspondem a abstrações simplificadas de organismos reais. Usando o computador é possível evoluir sistemas vivos virtuais, como redes gênicas e protéicas artificiais, e estudar seu desenvolvimento sob condições desejadas. Dessa forma, os atributos relevantes do sistema podem ser facilmente manipulados para serem propriamente adaptados aos propósitos da pesquisa, e as redes obtidas irão certamente apresentar configurações alternativas a cada nova execução do processo evolutivo.

Considerando esta possibilidade, é proposto aqui um modelo computacional conexionista de redes gênicas e protéicas, e tentamos evoluir sistemas artificiais que são capazes de realizar tarefas dinâmicas complexas. No modelo proposto, a rede é representada como um grafo direcionado, no qual nós correspondem a entidades biomoleculares e arcos são conexões funcionais representando reações bioquímicas descritas na forma de equações a diferenças. O modelo é utilizado em conjunto com uma abordagem evolutiva, em que, a partir de estruturas elementares, uma população de redes evolui através de mutações estruturais, considerando sua contínua interação com o ambiente.

Por conexionismo refiro-me à capacidade de processar informação e representar conhecimento de maneira distribuída, por meio de fluxo de informação quantitativa através de uma estrutura de rede interconectada composta de nós e conexões funcionais, assim como no formalismo de redes neurais artificiais (RNAs) (HAYKIN, 1994). Entretanto,

diferente das redes neurais tradicionais, o sistema de redes gênicas artificiais não é restrito a apenas um tipo de nó computacional (o clássico modelo de neurônio da literatura de RNAs), mas inclui diferentes funcionalidades não-lineares e lineares, determinadas por um conjunto de reações bioquímicas, arranjadas em uma estrutura assimétrica.

Essa abordagem se baseia na suposição de que: (*i*) redes gênicas e protéicas não podem ser completamente compreendidas através da decomposição de suas propriedades em unidades menores (isto é, a partir de um ponto de vista puramente reducionista)[4] (PRIGOGINE & STENGERS, 1984; KAUFFMAN, 1993); (*ii*) sistemas vivos são sistemas abertos, que evoluem em permanente interação dinâmica com um ambiente, e eles devem ser estudados sob uma perspectiva integrativa (SCHNEIDER & KAY, 1994); e (*iii*) as propriedades estruturais das redes celulares são determinantes para o seu funcionamento e dinâmica (KAUFFMAN, 1993; STROGATZ, 2003). Conseqüentemente, dinâmica não-linear e arquitetura de rede não podem ser isoladas uma da outra.

Como ilustração da aplicabilidade do modelo, é estudado o caso particular em que configurações de rede alternativas são evoluídas para resolver um problema clássico de robótica autônoma modelado como uma tarefa de quimiotaxia. O agente, neste caso uma bactéria virtual, luta pela sobrevivência interagindo dinamicamente com o ambiente. Nesse problema multi-objetivo, a bactéria deve ser capaz de evitar toxinas mortais enquanto maximiza o consumo de nutrientes.

Embora o modelo proposto seja aplicado aqui como técnica de solução de problemas, o foco principal é dado à capacidade do sistema em representar e explicar características observadas em organismos reais. Como será demonstrado, as redes artificiais representam uma oportunidade *in silico* promissora para investigar a relação entre estrutura, dinâmica e comportamento em redes gênicas.

O restante deste capítulo está organizado da seguinte forma. Na Seção 3.3, uma breve revisão da literatura em evolução de redes gênicas artificiais é apresentada, e as características das abordagens existentes são contrastadas com as do modelo conexionista proposto. A Seção 3.4 descreve o modelo e a forma como ele é implementado. A Seção 3.5

[4] Esse ponto está relacionado a um debate polêmico entre as perspectivas reducionista e holística. A idéia central é que a emergência de funcionalidade e propriedades de alto nível em um sistema complexo seria resultado do conjunto como um todo apenas, não podendo ser decomposta em partes menores. Tal conjuntura representaria uma limitação à perspectiva reducionista. Veja CAPRA (1982) para uma discussão sobre o tema.

define a modelagem do problema de quimiotaxia, e a Seção 3.6 descreve o procedimento evolutivo empregado. Os experimentos computacionais e os seus resultados são apresentados na Seção 3.7. Na Seção 3.8, o modelo conexionista é visto como uma técnica de solução de problemas, e um paralelo é traçado entre as redes gênicas artificiais e as redes neurais artificiais, a abordagem conexionista mais tradicional. A Seção 3.9 conclui o capítulo, trazendo uma discussão geral sobre as conclusões extraídas dos experimentos.

3.3. Revisão da literatura: evolução de redes gênicas *in silico*

O projeto evolutivo de redes gênicas *in silico* é uma tendência crescente nos estudos de sistemas genéticos reguladores, mas há ainda relativamente poucos trabalhos propostos na literatura cobrindo este tópico. Nesta seção, alguns dos sistemas genéticos reguladores artificiais existentes na literatura serão revisados e suas características contrastadas com as do modelo conexionista apresentado neste capítulo.

REIL (1999) foi o primeiro a sugerir o projeto evolutivo de redes reguladoras artificiais. Ele propôs o Genoma Artificial (*Artificial Genome* – AG), uma extensão do trabalho pioneiro de Kauffman em dinâmica de redes booleanas aleatórias (KAUFFMAN, 1993), que incorpora interações reguladoras mais plausíveis biologicamente. O genoma é representado em forma de uma string de inteiros variando de 0 a 3 (correspondendo aos quatro tipos de nucleotídeos) e as interações reguladoras são determinadas por casamento de strings.

HALLINAN & WILES (2004a) aplicaram um algoritmo evolutivo para busca de genomas artificiais que apresentam dinâmica de ciclo limite, e estudaram a influência de atualização síncrona e assíncrona na dinâmica do modelo (HALLINAN & WILES, 2004b). Aspectos estruturais da rede foram analisados apenas em termos de grau de conectividade.

Redes booleanas são uma abordagem interessante para modelagem de redes reguladoras, dado seu potencial para exibir dinâmica complexa – como caos e ciclo limite – e também devido ao seu reduzido custo computacional, o que torna possível simular redes de grande porte. Entretanto, sua natureza discreta binária é uma simplificação muito grande e as conclusões produzidas por esse modelo excessivamente abstrato dificilmente podem ser generalizadas. Além disso, as funcionalidades adotadas no modelo de REIL (1999) são restritas apenas à síntese de proteínas reguladoras e regulação gênica direta (não existem

interações proteína-proteína), o sistema é considerado isolado, isto é, não há interações sistema/ambiente, e as redes não realizam tarefa alguma, não havendo, portanto, processamento útil de informação.

BONGARD (2002) propõe um sistema intrincado chamado ontogenia artificial (*Artificial Ontogeny* – AO). Esse sistema é baseado na combinação de um modelo de genoma artificial e redes neurais, e é usado na evolução de comportamento motor em robôs virtuais. Embora os robôs interajam com o ambiente, o sistema de ontogenia artificial é muito específico ao problema, e não apresenta papel relevante no entendimento do funcionamento dos sistemas reguladores.

KUO *et al*. (2004) propuseram um modelo de redes reguladoras artificiais (*Artificial Regulatory Networks*) baseado em equações diferenciais, que é utilizado em conjunto com um procedimento evolutivo para reproduzir funções trigonométricas bastante simples, como a função seno. O genoma e as proteínas são codificados na forma de strings binárias, e as interações reguladoras são determinadas através de casamento de strings. De forma semelhante ao genoma artificial de REIL (1999), as únicas reações consideradas são regulação gênica e síntese de proteína, e o sistema é fechado à informação externa. Os autores não consideram a análise da estrutura de suas redes.

FRANÇOIS & HAKIM (2004) desenvolveram um interessante sistema evolutivo baseado em modelagem tradicional de redes gênicas com equações diferenciais ordinárias (DE JONG, 2002). O modelo evolui através da adição de novas equações e pela mutação em seus parâmetros cinéticos; o sistema de equações resultante é resolvido utilizando integração numérica. As equações correspondem a reações bioquímicas, que nesse caso incluem não só regulação gênica e síntese de proteínas, mas também dimerização, fosforilação, entre outros (ou seja, interações proteína-proteína). O sistema é considerado isolado.

Diferente das redes booleanas, o modelo baseado em equações não dá ênfase à estrutura, mas às características idiossincráticas das reações envolvidas. Ademais, sua estrutura de dados é difícil de manipular em computador, conduzindo a um sistema evolutivo bastante inflexível. O modelo é empregado na implementação de uma chave bi-estável e de um oscilador permanente (FRANÇOIS & HAKIM, 2004). Essa mesma estratégia é

adotada por DECKARD & SAURO (2004) para evoluir soluções analógicas para operações aritméticas simples.

Levando em consideração os aspectos positivos e negativos das redes artificiais descritas aqui, é possível enfatizar que:

- O modelo conexionista a ser proposto neste capítulo é estrutural em essência. O foco é dado às propriedades estruturais da rede (como topologia, grau de conectividade, coeficiente de clusterização e força das conexões) e sua influência na dinâmica;
- Nosso modelo inclui um rico repertório de reações bioquímicas e funcionalidades, sendo capaz de reproduzir dinâmica não-linear complexa, e pode ser facilmente estendido para incorporar novas reações e novos componentes;
- A representação conexionista é flexível e simples de ser implementada. O modelo permite fácil manipulação de sua estrutura de dados;
- O sistema é inerentemente aberto e interativo, considerando assim as relações integrativas[5] com o ambiente, uma característica essencial dos sistemas vivos. A informação do ambiente é codificada na forma de entidades biomoleculares, e o sistema é exposto à variação na concentração de moléculas, assim como uma célula é exposta à variação na concentração de compostos químicos em sua vizinhança.

3.4. O Modelo Conexionista

A. Representação

O modelo conexionista consiste em um grafo direcionado, no qual nós correspondem a diferentes tipos de moléculas, como proteínas e genes, e os arcos estão associados a relações matemáticas entre esses elementos.

Essa estrutura é implementada na forma de uma matriz quadrada de conectividade, na qual cada linha/coluna representa um dado nó da rede e os elementos da matriz diferentes de zero correspondem aos arcos. Um arco na *i*-ésima linha representa as conexões que saem do nó *i*, enquanto os arcos na *i*-ésima coluna, as conexões que chegam

[5] O termo "relações integrativas" é utilizado no sentido de considerar o sistema como sendo integrado ao ambiente, de acordo com uma perspectiva sistêmica.

ao nó *i*. Adicionar um novo nó à rede leva à introdução de uma linha e de uma coluna adicional na matriz. Para remover um nó, basta remover a linha e coluna da matriz que o representa, e todos os arcos que relacionam este nó com o resto da rede serão, dessa forma, removidos também.

Os nós podem ser de seis tipos diferentes, como resumido na Tabela 3.1. Apenas 6 tipos de nós foram considerados nessa implementação, mas tipos adicionais podem ser incorporados. Note que as proteínas de entrada na Tabela 3.1 são consideradas como nós do tipo II para fins de implementação. Essa distinção entre nós é necessária, pois tipos diferentes de nós participam em reações diferentes, ou podem assumir papéis diferentes em uma mesma reação.

Tabela 3.1 Possíveis tipos de nós do modelo e o esquema de cores utilizado para representá-los.

Tipo		Descrição
I	gene	Um gene, o qual sempre está associado à produção de uma proteína. O gene pode se ligar a uma proteína reguladora.
II	produto do gene	Uma proteína que é o produto direto de um dado gene.
III	dímero	Um homo ou heterodímero, formado pela junção de duas proteínas.
IV	complexo gene-proteína	Um nó que representa a junção de uma proteína reguladora a um gene.
V	proteína fosforilada	Uma proteína após ser fosforilada em uma reação enzimática.
VI	complexo enzima-substrato	Um nó representando a junção de uma enzima (proteína) a um substrato (outra proteína).
Entrada	proteína de entrada	Representa as variáveis do ambiente (toxinas e nutrientes). Para os propósitos de implementação das reações descritas nesta seção, as proteínas de entrada serão consideradas como do tipo II.

A síntese de uma rede está associada a um processo evolutivo. Nesse processo, a rede começa com uma estrutura elementar, digamos, um gene, e cresce através da adição de reações, como dimerização e fosforilação protéica, assim como em FRANÇOIS & HAKIM (2004). Um nó nulo é também necessário para representar a degradação das proteínas (veja a Figura 3.1). A escolha das reações e dos seus parâmetros é realizada aleatoriamente (são mutações estruturais), mas, obviamente, o processo evolutivo será responsável por selecionar as estruturas que produzem o efeito desejado.

As reações consideradas aqui e a forma com que elas são representadas na rede são mostradas na Figura 3.1, e podem ser descritas como segue:

1) Adicionar gene – Figura 3.1 (a):

   Descrição: Esta reação consiste simplesmente na síntese de proteína a partir de um gene. O gene possui uma taxa de produção de proteína fixa, mas essa taxa pode ser alterada por meio de proteínas reguladoras.

   Implementação: Para inserir um gene, adicione dois nós à rede – um gene (tipo I) e a proteína que o gene produz (tipo II) – e implemente as ligações mostradas na figura. Os arcos na figura correspondem à síntese de proteína por um gene, e a degradação dessa proteína. A funcionalidade dos arcos será descrita mais adiante, na Seção 3.4.B.

2) Adicionar dimerização – Figura 3.1 (b):

   Descrição: Dimerização é a formação de um complexo protéico através da junção de duas proteínas. A nova proteína formada pode ter propriedades e funcionalidades completamente diferentes das proteínas individuais que a formaram. A dimerização está envolvida em vários processos reguladores e de sinalização.

   Implementação: Selecione aleatoriamente duas proteínas *A* e *B* (tipos II, III ou V), adicione um novo nó para o dímero (tipo III) e implemente as ligações mostradas na Figura 3.1 (b). A mesma proteína pode ser selecionada duas vezes, dando origem a um homodímero.

3) Adicionar proteína reguladora – Figura 3.1 (c):

Descrição: Uma proteína reguladora é uma proteína capaz de se ligar a um gene (ou, mais especificamente, ao promotor do gene) modulando a sua expressão. Um mesmo gene pode sofrer influência de várias proteínas reguladoras, que podem atuar de forma competitiva ou cooperativa. Aqui, apenas a regulação competitiva é considerada.

Implementação: Essa reação adiciona uma nova proteína reguladora *B* (tipos II, III ou V) ao gene *a* (tipo I). Selecione aleatoriamente *B* e *a*, crie um novo nó *aB* (tipo IV) e implemente as conexões mostradas na figura.

4) Adicionar fosforilação enzimática – Figura 3.1 (d):

Descrição: Fosforilação consiste na adição de um ou mais grupos fosfato (PO_4) a uma determinada proteína, um processo geralmente catalisado por uma enzima. A fosforilação pode alterar completamente as propriedades e funcionalidades de uma proteína. Este mecanismo, juntamente com a desfosforilação, é provavelmente o evento regulador mais importante em eucariotos.

Implementação: Essa reação requer a inserção de dois nós. Selecione duas proteínas *A* e *E* aleatoriamente (tipos II, III ou V) e crie um nó *EA* enzima-substrato (tipo VI). A seguir, adicione uma proteína fosforilada *A** (tipo V) na rede e implemente as conexões ilustradas na figura.

5) Adicionar degradação enzimática – Figura 3.1 (e):

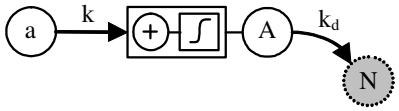
Descrição: Degradação enzimática, como considerada aqui, consiste na decomposição parcial de um complexo protéico por meio da atuação de uma enzima.

Implementação: Selecione aleatoriamente um dímero *AB* (tipo III) e outra proteína qualquer *E* (tipos II, III ou V) e insira um novo nó enzima-substrato *EAB* (tipo VI). Implemente as conexões mostradas na Figura 3.1 (e).

(a) **Adição de gene**

$a \xrightarrow{k} a + A$

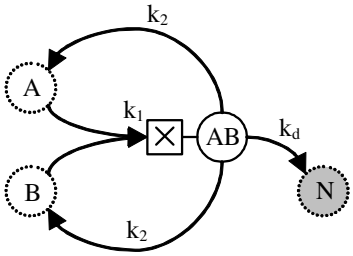
$A \xrightarrow{k_d} N$



(b) **Adição dimerização**

$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} AB$

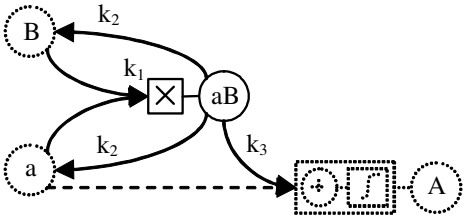
$AB \xrightarrow{k_d} N$



(c) **Adição de ação reguladora**

$a + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} aB$

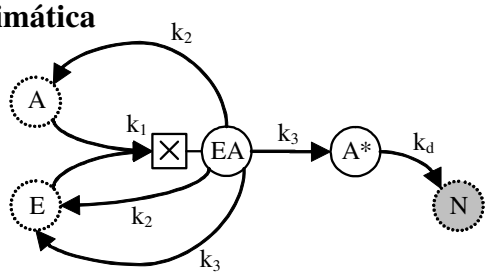
$aB \xrightarrow{k_3} aB + A$



(d) **Adição de fosforilação enzimática**

$A + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} EA \xrightarrow{k_3} E + A*$

$A* \xrightarrow{k_d} N$



(e) **Adição de degradação enzimática**

$AB + E \underset{k_2}{\overset{k_1}{\rightleftharpoons}} EAB \xrightarrow{k_3} E + A$

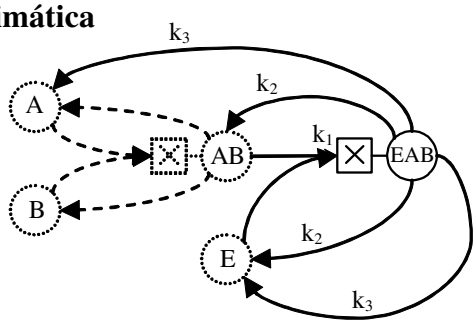


Figura 3.1 Reações e seus correspondentes diagramas conexionistas. Círculos representam os nós da rede, onde letras minúsculas são genes e letras maiúsculas, proteínas. O nó cinza "*N*" é o nó nulo, que representa o

destino das proteínas degradadas. k_i (i = 1,2,3) representam as constantes cinéticas e k_d é a constante de degradação. As caixas quadradas e retangulares próximas a alguns dos nós representam a funcionalidade associada às conexões, descritas mais adiante na Figura 3.2. Círculos e linhas tracejadas correspondem a nós e conexões já existentes, enquanto as linhas contínuas representam nós e conexões que devem ser adicionados em cada reação. (a) Adição de gene. (b) Adição de dimerização. (c) Adição de ação reguladora. (d) Adição de fosforilação enzimática. (e) Adição de degradação enzimática.

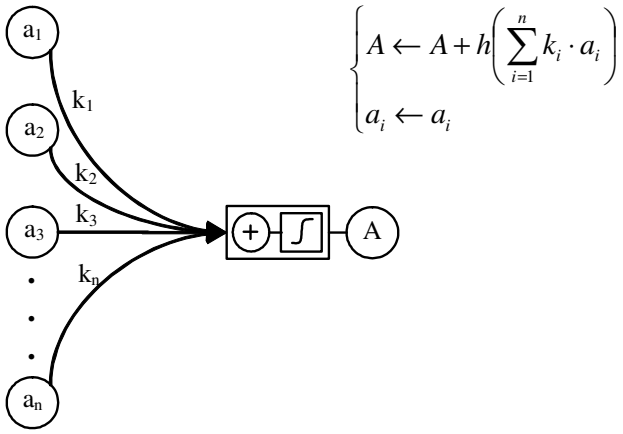
Essas 5 reações fornecem um rico repertório de configurações estruturais e provêm o básico em flexibilidade e operações não-lineares, embora o modelo não esteja de forma alguma completo. Muitas outras reações podem ser modeladas e incluídas no sistema. Um bom exemplo é a desfosforilação, que é sabido estar envolvida em vários processos reguladores em organismos eucariotos. Em teoria, quanto mais reações são incorporadas, mais flexibilidade é adquirida para ser propriamente explorada pelo usuário.

Embora muitas das reações da Figura 3.1 tenham sido empregadas também em FRANÇOIS & HAKIM (2004), é importante distinguir que elas são modeladas aqui de maneira mais completa. O processo de síntese de proteína na reação de adição de gene mostrada na Figura 3.1 (a) inclui um operador não-linear envolvendo a função *hill curve* (como explicado na próxima subseção), que é ignorado em FRANÇOIS & HAKIM (2004). Além do mais, as enzimas nas reações de fosforilação enzimática e degradação enzimática, modeladas aqui como variáveis do sistema, são consideradas como sendo constantes em FRANÇOIS & HAKIM (2004), o que transforma as duas reações em operações puramente lineares.

B. Simulação

O modelo é simulado por meio de propagação em tempo discreto. Cada nó da rede é uma variável do sistema, e o estado da variável denota a concentração de seu tipo molecular. A cada unidade de tempo, o estado das variáveis é atualizado baseado no último estado, de acordo com as conexões da rede. No início da simulação, todas as variáveis assumem valor zero, com exceção dos genes, que assumem valor 1. O fluxo na rede é calculado através de três tipos diferentes de conexões funcionais, como descrito na Figura 3.2.

(a)



(b)

(c)

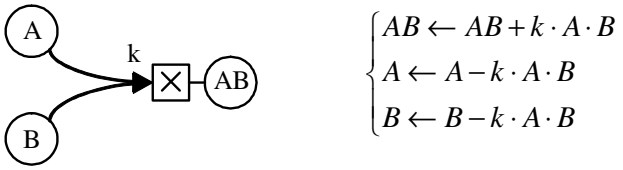


Figura 3.2 Descrição das conexões funcionais da rede. (a) Síntese de proteína: esse tipo de conexão é um regressor não-linear; ela primeiramente soma a contribuição de todas as entidades reguladoras, e essa soma determina não-linearmente a produção da proteína *A*, de acordo com a função *hill curve h*. (b) Conexão linear: descreve o fluxo linear de uma proteína *A* para outra, *B*. (c) Conexão produto: descreve a junção de duas moléculas para formar uma terceira.

A Figura 3.2(a) ilustra a atualização dos estados em um processo de síntese de proteína. Várias proteínas reguladoras podem se ligar a um mesmo gene, formando os diferentes nós a_i na figura. Cada um desses nós do tipo IV, juntamente com o nó original, tipo I, vai dar sua própria contribuição para a síntese da proteína *A*. Note que, quando k_i é grande, a proteína reguladora trabalha estimulando a ativação do gene e quando k_i é

pequeno, ela tende a suprimir a atividade do gene. Essas contribuições são somadas e usadas como parâmetro para a função *hill curve*, a qual determina a quantidade de proteína a ser produzida. A função *hill curve* é muito utilizada para modelar o controle regulador de forma biologicamente plausível (DE JONG, 2002). Ela tem uma conformação sigmoidal e pode ser descrita analiticamente pela equação 3.1:

$$h(s,\theta,m) = \frac{s^m}{s^m + \theta^m}\ , \tag{3.1}$$

onde s é a influência reguladora total, θ é um limiar para a influência reguladora e m é uma constante que determina a inclinação da curva. Nesse trabalho, θ e m foram configurados em 0,5 e 3, respectivamente.

É interessante notar que, apesar das particularidades envolvidas aqui (como a limitação na concentração dos genes ou o fato de as concentrações não assumirem valores negativos), a estrutura na Figura 3.2(a) é essencialmente a mesma de um modelo de neurônio da literatura de redes neurais artificiais (HAYKIN, 1994). Ou seja, um gene, modelado da forma apresentada aqui, executa o papel de um regressor múltiplo não-linear. Redes compostas somente dessas unidades estruturais são capazes de realizar mapeamentos complexos do tipo entrada-saída. Essa estrutura consiste em uma poderosa e flexível ferramenta computacional.

A Figura 3.2(b) ilustra as equações de atualização para conexões lineares, e a Figura 3.2(c), as equações de atualização para propagação em produto, envolvidas em dimerizações.

3.5. Modelagem do Problema de Quimiotaxia

Quimiotaxia é a capacidade de um organismo em guiar-se baseado no gradiente de concentração de compostos químicos em um ambiente. No problema considerado aqui, uma bactéria deve ser capaz de evitar elementos repelentes (tóxicos) e se dirigir para regiões de alta concentração de elementos atratores (nutrientes).

Na modelagem empregada, bactérias virtuais são pontos móveis, com velocidade constante, em um ambiente bidimensional numa região compacta [0,1]×[0,1]. Para cada bactéria, uma rede gênica diferente está associada. O ambiente contém elementos tóxicos e

também nutrientes, cujas concentrações são modeladas por distribuições gaussianas. Essas distribuições determinam a quantidade de toxinas e de nutrientes à qual uma bactéria localizada em uma determinada posição do espaço está exposta. A Figura 3.3 mostra uma imagem do ambiente que será utilizado no problema. Note que há regiões de sobreposição de áreas tóxicas e de nutrientes, gerando nessas regiões objetivos concorrentes.

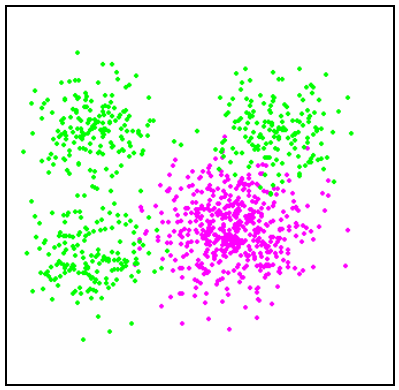


Figura 3.3 Ambiente bidimensional onde as bactérias viverão a cada geração. Pontos em verde ilustram a concentração de toxinas; pontos em rosa ilustram a concentração de nutrientes. Veja que há três focos de concentrações de toxinas e apenas um de nutrientes. Todos seguem distribuições gaussianas.

Uma bactéria elementar possui uma rede simples composta de 5 nós: um nó nulo, dois nós que representam as proteínas de entrada (representando a toxina e os nutrientes, que variam em estado de acordo com a posição da bactéria), um gene e a proteína que este gene produz. Essa última proteína é selecionada para ser o atuador. A bactéria tipicamente nada em linha reta e se em algum momento a concentração do atuador cai, ela realiza uma curva para um lado (direito ou esquerdo), escolhido aleatoriamente. O ângulo da curva depende da variação na concentração da proteína, de acordo com a equação 3.2:

$$\alpha = \alpha + rand \times \tan^{-1}(\Delta A), \tag{3.2}$$

onde α é o ângulo do vetor de trajetória da bactéria, *rand* é um número inteiro que pode assumir ±1 e ΔA é a queda na concentração da proteína atuadora. A função $\tan^{-1}$ empregada aqui é interessante por permitir uma variação máxima de ±90 graus de mudança na direção de movimentação da bactéria, mesmo para um ΔA muito grande. Uma vez que a bactéria começa uma curva, ela vai sempre virar para o mesmo lado (*rand* permanece constante) até que a concentração de *A* pare de cair. Note que, inicialmente, não há conexões entre as

proteínas do ambiente e o atuador; as bactérias literalmente ignoram a informação do ambiente. É esperado que, pela adição de reações aleatórias e pressão seletiva, as redes gênicas serão capazes de mapear de alguma forma a informação de entrada em um comportamento de saída, modulado pelas variações na concentração da proteína atuadora.

3.6. Procedimento Evolutivo

O procedimento evolutivo começa com uma população inicial de 40 bactérias elementares. Embora populações menores com 20 ou até 10 indivíduos também tenham sido capazes de evoluir o comportamento de quimiotaxia, uma população de 40 indivíduos é mais eficiente, e mostrou-se capaz de resolver o problema em praticamente todas as execuções do processo evolutivo. Para cada bactéria da população, uma das reações da Figura 3.1 é aleatoriamente selecionada e adicionada, e a população inteira é colocada pra interagir com o ambiente por 300 iterações – uma iteração corresponde a um passo de tamanho 0,02 no ambiente bidimensional de tamanho 1×1. O número de iterações para avaliação e o tamanho do passo foram determinados empiricamente. O número de iterações deve ser grande o suficiente para permitir que a bactéria encontre áreas tóxicas e de nutrientes em sua trajetória, mas pequeno o suficiente para limitar o custo computacional da avaliação. O tamanho do passo deve ser pequeno para dar à bactéria tempo suficiente para propriamente perceber e reagir às entradas, mas passos muito pequenos não são convenientes, pois o número de iterações necessário teria de ser grande demais.

A performance de uma bactéria é avaliada de acordo com a quantidade de nutrientes que ela acumulou, e uma penalidade é atribuída à complexidade de sua rede. Complexidade aqui é considerada como o tamanho da rede gênica, isto é, o número de nós que ela contém. A equação 3.3 descreve o cálculo do *fitness*, o qual é baseado em um mecanismo de ranking:

$$fit = rank_f - \frac{rank_c}{5}, \tag{3.3}$$

onde $rank_f$ é o ranking da bactéria relativo à quantidade de nutrientes acumulados e $rank_c$ é o ranking relativo à sua complexidade. A penalização da complexidade é necessária para controlar o crescimento das redes, evitando assim a evolução de redes muito grandes e com muitos nós sem utilidade efetiva. Note que as toxinas não fazem parte diretamente da

função de *fitness*. Entretanto, se uma bactéria atinge uma concentração crítica de toxinas, ela morre, e é eliminada da etapa de seleção.

Baseado no maior *fitness*, 8 dentre as 40 bactérias são selecionadas para a próxima geração. Cada uma delas produz 4 cópias mutadas de si mesma, compondo os 32 indivíduos remanescentes da população. A seguir, o procedimento começa novamente para essa nova geração de bactérias e é repedido por 20 gerações. Esse algoritmo consiste num procedimento evolutivo elitista bastante simples. Nenhuma sofisticação relativa ao processo de seleção é adotada aqui, e também nenhum operador de *crossover* é empregado.

Mutações consistem em remover ou adicionar uma das cinco reações descritas na Figura 3.1. Quando uma reação é adicionada, as constantes cinéticas k_i são determinadas por valores aleatórios entre 0 e 1. A constante de degradação foi arbitrariamente configurada para 0,1 em todos os casos. Adição de reações tem probabilidade 0,7 e a probabilidade de remoção é 0,3. Note que a mutação é essencialmente estrutural. A rede evolui pela adição e remoção de nós e conexões apenas, e ajuste nos parâmetros não é permitido.

Como não há informação disponível sobre taxas de mutação estrutural na natureza, os parâmetros de mutação empregados são arbitrários, e o mesmo foi feito para a constante de degradação. Entretanto, é importante salientar que o procedimento evolutivo se mostrou robusto aos parâmetros. Embora os valores usados aqui tenham sido determinados empiricamente para uma performance otimizada, o sistema vai funcionar para diferentes configurações paramétricas. Se, por exemplo, a probabilidade de adição de reação for mudada para 0,3 e a de remoção para 0,7 (isto é, os valores forem invertidos), uma boa solução para o problema de quimiotaxia também será obtida, embora um número maior de gerações vai ser necessário.

Um detalhe adicional diz respeito ao modo com que a adição de um gene é realizada. Ao invés de simplesmente adicionar um gene, como descrito na reação da Figura 3.1 (a), é realizada uma duplicação gênica, na qual um gene existente é escolhido aleatoriamente para ser duplicado. Nesse processo, 50% das iterações imediatas do gene antigo são herdadas pelo novo gene. Duplicação gênica é uma maneira mais realista de aumentar o tamanho da rede, e é um dos mecanismos responsáveis por gerar os padrões fractais observados em redes de organismos naturais (HALLINAN, 2004).

3.7. Experimentos

Os experimentos foram realizados de forma a testar a capacidade do sistema em evoluir o comportamento de quimiotaxia. A Figura 3.4 mostra a evolução de uma população de bactérias, ilustrando seu estado no ambiente após 300 iterações para 1, 5, 10 e 20 gerações. Na modelagem adotada, as bactérias não podem se interceptar. Elas agem individualmente sem serem afetadas pela presença umas das outras.

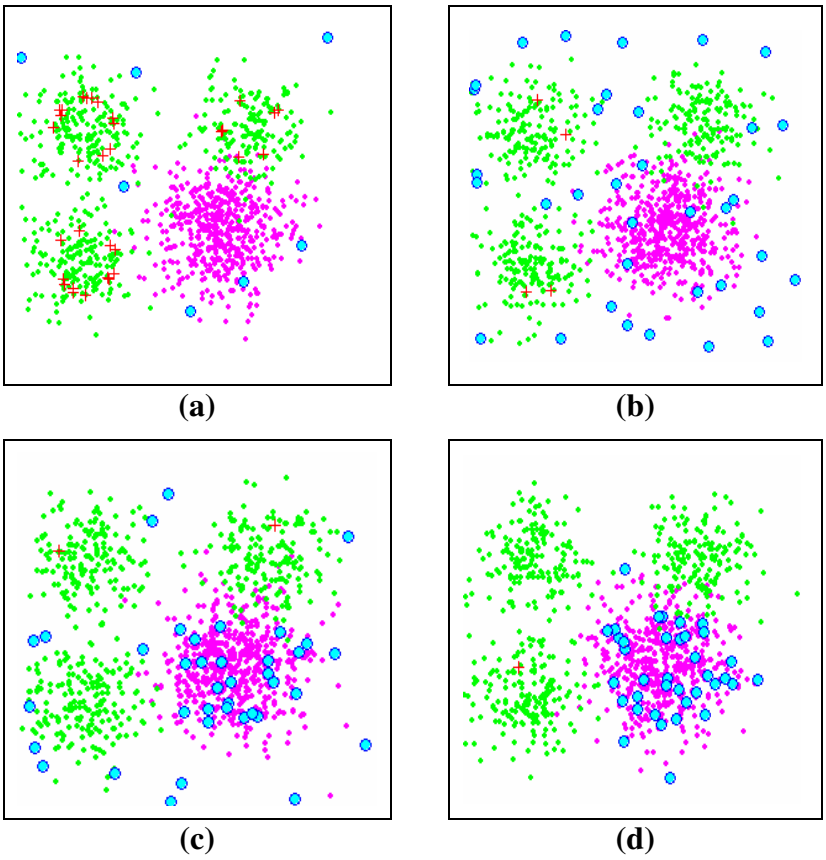


Figura 3.4 Posição das bactérias após 300 iterações. Cada gráfico representa um retrato instantâneo, pois as bactérias se encontram em movimento permanente. Círculos azuis representam bactérias vivas e cruzes vermelhas, bactérias mortas; pontos em verde ilustram a concentração de toxinas; pontos em rosa ilustram a concentração de nutrientes. (a) Após 1 geração. (b) Após 5 gerações. (c) Após 10 gerações. (d) Após 20 gerações.

Na primeira geração, uma porção significativa da população de bactérias morre por atingir níveis intoleráveis de concentração de toxinas (Figura 3.4 (a)). No entanto, algumas delas se mostraram capazes de evitar as zonas tóxicas e sobreviver. Como apenas 7 bactérias sobrevivem nesta primeira geração, todas elas (mais especificamente, as redes

gênicas associadas a elas) são selecionadas para a próxima etapa, de acordo com o processo evolutivo. Após 5 gerações de mutação e forte pressão seletiva (pois a morte de bactérias está envolvida), quase todas as bactérias podem evitar as toxinas e, assim, permanecerem vivas, como ilustrado na Figura 3.4 (b), mas elas ainda não se mostraram atraídas pelos nutrientes. Na geração de número 10, no entanto, algumas bactérias parecem já ter desenvolvido o comportamento de consumo de nutrientes, e dão preferência a permanecer sobre as áreas de alta concentração de nutrientes, em vez de vagar aleatoriamente pelo ambiente (Figura 3.4 (c)). Após 20 gerações, este comportamento foi disseminado pela população resultante, e quase todas as bactérias preferem permanecer sobre as regiões de alta concentração de nutrientes, enquanto continuam sendo capazes de evitar as regiões tóxicas (Figura 3.4 (d)).

A. Análise da estrutura

A Figura 3.5 (a) mostra a rede do melhor indivíduo da população para esta execução em particular do processo evolutivo. O esquema de cores utilizado é o descrito na Tabela 3.1. A estrutura é mostrada em um panorama conexionista, mas, diferente dos diagramas da Figura 3.1, os detalhes das conexões são omitidos de forma a enfatizar a topologia.

Veja que apenas um gene é necessário para resolver esse problema, embora configurações com mais de um gene também podem aparecer na população. O atuador é o nó com mais conexões, um total de 7, e seu gene possui 3 interações reguladoras. A rede possui ainda 3 dimerizações e uma reação enzimática.

Embora nada de muito relevante possa ser inferido diretamente pela simples inspeção da estrutura estática da rede, uma análise da dinâmica das variáveis pode revelar muito sobre o funcionamento do sistema. As Figura 3.5 (b), (c) e (d) mostram a evolução dos estados das variáveis quando a bactéria se aproxima da região tóxica. Inicialmente, não há toxinas próximo à bactéria, e o nó 1, que representa essa informação do ambiente, está completamente branco (Figura 3.5 (b)). A seguir, a bactéria se aproxima da região tóxica e os nós 5 e 6 começam a escurecer (Figura 3.5 (c)). A bactéria se aproxima ainda mais do centro da dispersão de toxinas, como ilustrado pela cor do nó 1 na Figura 3.5 (d). Agora, vários nós estão ativados e a concentração da variável 4 finalmente cai, fazendo a bactéria virar.

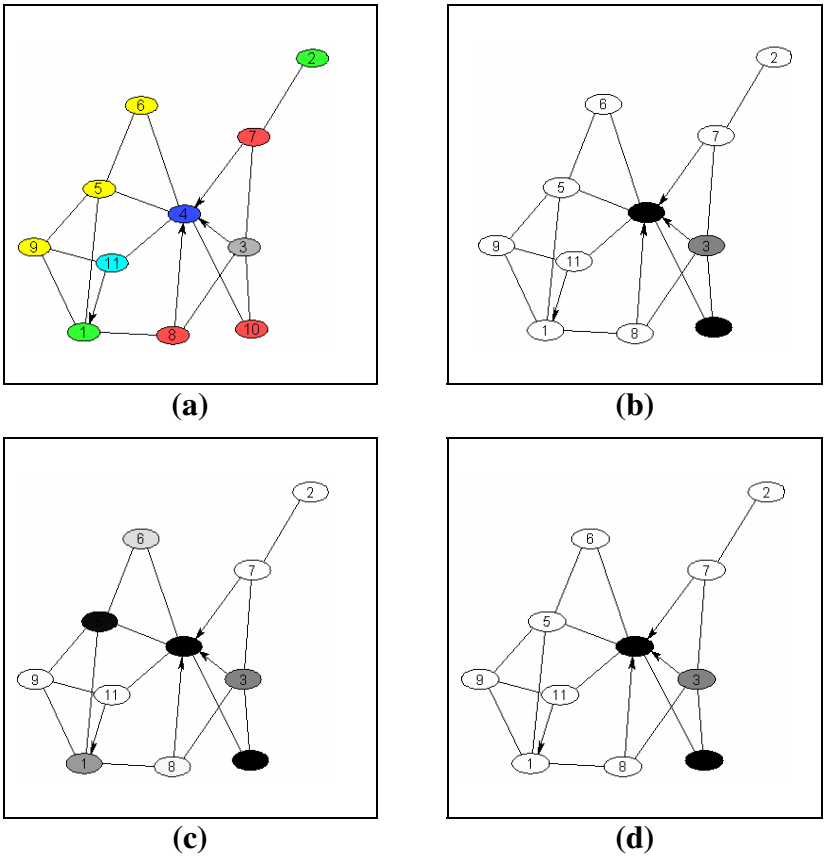


Figura 3.5 (a) Estrutura da rede evoluída; o nó nulo não é mostrado. Os números dos nós indicam a ordem em que eles apareceram durante o processo evolutivo. Ligações não direcionadas denotam a presença de conexões de ida e de volta entre os nós. O nó 1 representa a informação de concentração de toxina e o nó 2, a informação de concentração de nutrientes. O nó 4 é o atuador, que é sintetizado pelo gene do nó 3. (b) Estado das variáveis antes de se aproximar da zona tóxica. Cinza escuro significa alta concentração e cinza claro, baixa concentração. (c) Estado das variáveis quando a bactéria está próxima da zona tóxica. (d) Estado das variáveis quando a bactéria muda de trajetória.

Essa atividade dinâmica sugere que há muitos nós envolvidos no comportamento de evitar toxinas, e esse, de fato, parece ser o caso aqui. Para verificar essa hipótese, faremos uso do fato de que o sistema é virtual e literalmente retiramos alguns nós da rede para ver o impacto no comportamento. Comecemos pelos nós 9 e 11, que são menos ativados e, por conseguinte, parecem ter uma influência menor. De fato, quando ambos os nós são removidos, aparentemente nada diferente acontece, e a bactéria ainda mantém seu comportamento original. Entretanto, para o caso particular em que o sistema acaba de ser inicializado e a concentração da proteína atuadora ainda não estabilizou, esses nós são de

vital importância. Se, nas primeiras iterações, a bactéria se dirige diretamente para a região tóxica, ambos os nós são necessários para que a curva seja realizada a tempo e a toxina seja evitada. Se eles não estão presentes, o papel realizado pelos outros nós sozinhos não é suficiente, e a bactéria morre.

Quando o nó 6 é removido, o efeito é mais drástico. A curva realizada pela bactéria se torna visivelmente mais lenta e fraca, e ela não pode evitar todos os encontros com a zona tóxica, embora na maioria das vezes ela ainda consiga. Além disso, sem o nó 6 o efeito da toxina se torna persistente, fazendo a bactéria nadar erraticamente por algum tempo após o encontro com a zona tóxica, em vez de simplesmente seguir em linha reta. Se o nó 5 é removido, a habilidade da bactéria em evitar as zonas tóxicas é eliminada. Ela simplesmente ignora a presença de toxinas, e sempre morre nos encontros. O nó 8, embora fortemente ativado, não causou um alteração perceptível no comportamento quando removido, embora os testes não tenham sido exaustivos.

Essa análise confirma que a função é distribuída entre os nós da rede e o comportamento do sistema depende do conjunto inteiro de nós, e não de nós individuais. Não obstante, a remoção de um ou mais nós nem sempre é catastrófica e, na maioria das vezes, a rede ainda pode manter uma performance mínima. A mesma propriedade pode ser verificada para o comportamento de atração pelos nutrientes, o qual para esta rede é fortemente relacionado aos nós 7 e 10. Com ambos os nós, a bactéria é capaz de coletar uma quantidade média de 140 unidades de nutrientes em 300 iterações. Se o nó 10 é removido, essa quantidade cai aproximadamente para a metade, mas o comportamento de atração pelos nutrientes ainda é observado. Variações paramétricas nas conexões desses nós foram também realizadas e mostraram um efeito semelhante.

Essa característica interessante do sistema é uma conseqüência direta da maneira com que a rede cresce. Ao invés de simplesmente encontrar as conexões ótimas, com parâmetros bem ajustados que levam ao comportamento desejado, a rede usualmente começa com uma conexão não ótima que produz um comportamento imperfeito (nó 5, por exemplo). Em vez de optar pelo ajuste fino dos parâmetros das reações (o que, aliás, não é permitido aqui), o procedimento evolutivo adiciona mais nós à rede, que eventualmente assumem parte do trabalho e melhoram o comportamento do sistema. Como resultado, a

estrutura está tendo papel na otimização, de modo que tolerância a falhas e robustez paramétrica emergem naturalmente deste processo.

De fato, foi testado também um procedimento evolutivo que considera mutação paramétrica, em que não apenas adição e remoção de reações ocorre, mas otimização de seus parâmetros também é permitida. As redes resultantes são usualmente muito menores, apresentando reações otimizadas ao invés de um grupo de reações não-ótimas. Contudo, essas redes são mais sensíveis a variações paramétricas, e não apresentam robustez à remoção de nós. Foge ao escopo deste trabalho, no entanto, realizar investigações mais aprofundadas nesta linha.

B. Comportamento das bactérias

Para evitar as regiões tóxicas, a bactéria simplesmente muda de direção quando a concentração de toxinas aumenta. Este é um comportamento simples e é adquirido com facilidade pela rede gênica. Maximizar o consumo de nutrientes, no entanto, não é tão simples assim. A bactéria não pode parar; é forçada a se deslocar para sempre e o seu comportamento de "curva para um lado aleatório" modulado pelo atuador é muito limitado. A Figura 3.6 mostra a solução encontrada pelo processo evolutivo para esse problema. Quando a bactéria percebe a queda na concentração de nutrientes, ela faz uma curva. Essa curva é tão bem ajustada que a bactéria vai diretamente para o centro da dispersão, maximizando assim a quantidade de nutrientes absorvidos.

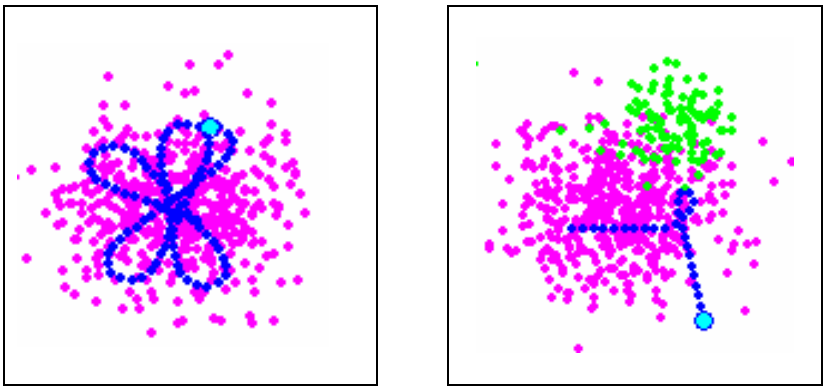


Figura 3.6 (a) Comportamento evoluído para a bactéria de maior *fitness*. (b) Situação de tomada de decisão.

Agora note que quando a bactéria está fazendo a curva de modo a retornar ao centro, ela vai diretamente contra o estímulo direto do ambiente, isto é, a concentração de nutrientes começa a aumentar, mas ela continua a curva. Ela não está simplesmente seguindo a regra "vire quando a concentração cair"; a bactéria parece ter desenvolvido o que se chama comportamento deliberativo (ARKIN, 1998). Usando sua dinâmica interna, ela incorpora o estímulo recebido e realiza a curva baseado na informação passada, e não na informação atual de seus sensores. Se os nutrientes são removidos do ambiente no momento da curva, a bactéria ainda mostra persistência e continua a curva inteira, retornando para onde o centro da dispersão de nutrientes estava localizado, antes de seguir em frente em linha reta. Conclui-se então que a dinâmica que rege a trajetória possui uma inércia ajustada à configuração ambiental definida.

Diferentemente disso, um agente puramente reativo, o qual não possui dinâmica interna, é guiado apenas pela informação instantânea de entrada. O melhor comportamento que um agente desse tipo poderia desenvolver é um circulo perfeito, representando um lugar geométrico de densidade praticamente constante, dado que a densidade decresce radialmente. Portanto, ele não seria capaz de se dirigir ao centro da distribuição.

Para mostrar que a rede não está simplesmente atrasando a informação de entrada, considere o caso em que toxinas são colocadas à frente da bactéria precisamente no momento da curva. A Figura 3.6 (b) mostra o resultado do experimento. Em vez de simplesmente ignorar a entrada corrente, a bactéria imediatamente muda de atitude e dá prioridade ao ato de evitar a toxina.

O comportamento complexo observado nas bactérias virtuais pode ser considerado como uma indicação do potencial da abordagem conexionista e evolutiva proposta. Persistência e capacidade de tomada de decisão baseada em informações passadas e correntes são características relacionadas à autonomia (BODEN, 1998), uma propriedade compartilhada por organismos vivos, uni e pluricelulares, e que está fortemente associada à cognição. Essas características emergem da não-linearidade inerente ao modelo, combinada com sua intrincada estrutura recorrente, o que resulta numa elaborada dinâmica de rede interna.

C. Estruturas alternativas

Nesta seção, é apresentada uma amostragem da diversidade das estruturas evoluídas para resolver o problema de quimiotaxia. A Figura 3.7 mostra quatro redes com características estruturais visivelmente distintas.

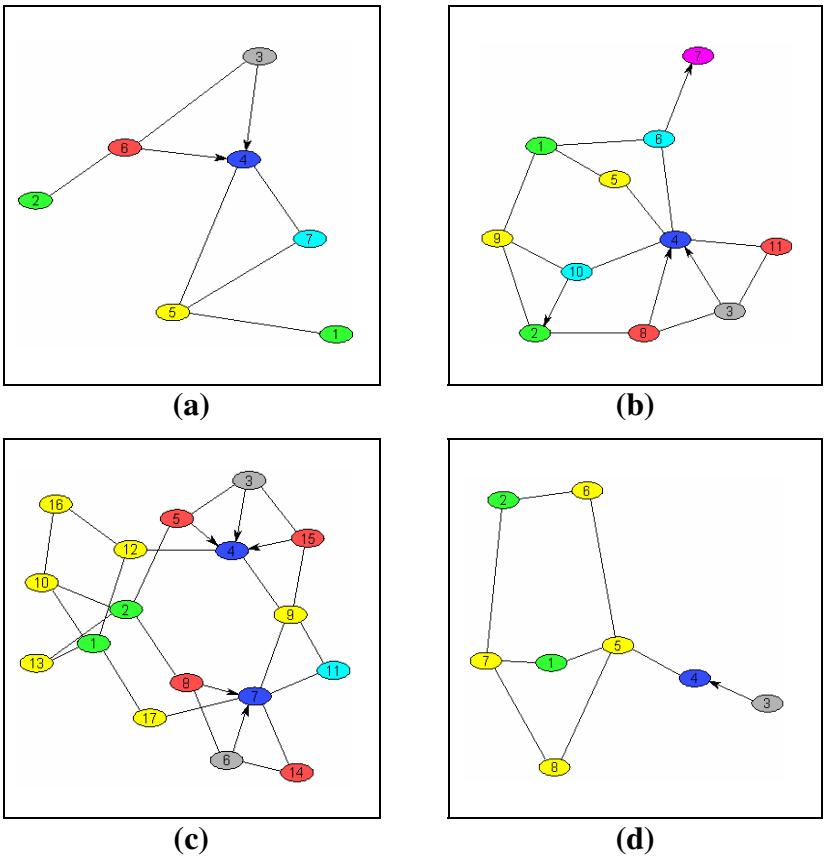


Figura 3.7 Diferentes configurações de rede evoluídas. O esquema de cores é o mesmo da Tabela 3.1. Assim como na Figura 3.5, apenas a topologia é enfatizada nesses diagramas.

A estrutura da Figura 3.7 (a) representa uma rede bastante otimizada. Ela possui poucos nós, embora apresente desempenho similar ao das outras estruturas. Entretanto, o seu funcionamento é fortemente dependente do ajuste de seus (relativamente poucos) parâmetros e, conseqüentemente, ela não é capaz de resistir a pequenas variações estruturais ou paramétricas. A rede da Figura 3.7 (b) é uma solução típica, considerando seu tamanho moderado, e a da Figura 3.7 (c) é uma rede maior, que possui dois genes. A função tende a ser mais distribuída para redes maiores, e elas, em geral, apresentam uma robustez a variações estruturais maior em relação às redes pequenas.

Uma análise da dinâmica das variáveis mostrou que essas três redes compartilham um princípio comum de funcionamento. Em todas elas, o comportamento de evitar toxinas é controlado principalmente pela redução direta na concentração do atuador, por meio de reações nas quais proteínas se ligam a ele, enquanto o comportamento de consumo de nutrientes é controlado através do processo de síntese, envolvendo, portanto, proteínas reguladoras. Entretanto, embora menos provável, outros tipos de configuração podem emergir. A Figura 3.7 (d) apresenta uma estrutura bastante distinta, nos termos considerados acima. Essa estrutura não possui proteína reguladora, e a atração por nutrientes é controlada pelo mesmo mecanismo empregado para evitar toxinas, isto é, vias compostas apenas por reações de dimerização e ausência de interações gene-proteína.

Esses quatro exemplos dão uma idéia da flexibilidade das redes gênicas e protéicas, reproduzidas aqui pelo modelo conexionista. A mesma tarefa pode ser realizada por várias configurações alternativas e as mesmas reações podem desempenhar papéis completamente diferentes em cada uma delas.

3.8. Redes Gênicas Artificiais

Usando o modelo conexionista de redes gênicas e de proteínas (será empregada aqui a nomenclatura redes gênicas artificiais – RGAs), foi possível resolver um problema clássico de robótica evolutiva (NOLFI & FLOREANO, 2002), um problema multi-objetivo bastante complexo, envolvendo dinâmica não-linear, aprendizado e adaptação. Os resultados sugerem a aplicação do modelo proposto como uma ferramenta computacional para resolução de problemas e processamento de informação, o que posicionaria as redes gênicas artificiais ao lado de abordagens conexionistas mais tradicionais, como as redes neurais artificiais.

No entanto, as RGAs apresentam várias características distintas em relação aos modelos tradicionais de redes neurais, que fazem delas uma classe particular de sistema conexionista. Os aspectos principais são resumidos a seguir:

- As RGAs são estruturais em essência. Como mostrado na análise da Seção 3.7.A, o conhecimento da rede está mais presente em sua topologia do que nos parâmetros. Como resultado, o sistema evoluído apresenta propriedades desejadas, como funcionalidade distribuída, tolerância a falhas e robustez paramétrica, que estão de

acordo com as características observadas em sistemas naturais. As redes neurais artificiais são em geral mais sensíveis à variação paramétrica, já que sua estrutura é predefinida e o conhecimento está representado nos parâmetros. Uma pequena modificação nos parâmetros de uma rede neural treinada pode alterar drasticamente seu comportamento, e a remoção de neurônios é geralmente intolerável em arquiteturas multicamadas convencionais. Essas são razões pelas quais a evolução dos parâmetros de redes neurais artificiais é uma tarefa complicada e exige etapas evolutivas mais elaboradas.

- Como conseqüência do paradigma estrutural, a configuração resultante não é definível a priori. Embora tenha sido arbitrada a escolha prévia do nó atuador, ele poderia simplesmente ter sido escolhido aleatoriamente (testes preliminares nesse sentido foram realizados com sucesso). A estrutura de redes neurais artificiais é em geral definida a priori, ou, quando evoluída, sua topologia é bastante restrita, e estruturas simétricas e conectividade completa ou em camadas são geralmente assumidas.
- Adicionalmente ao regressor não-linear (o modelo de neurônio das redes neurais), as RGAs possuem unidades não-lineares envolvendo inclusive operações de multiplicação na agregação de sinais e também a possibilidade de uma conexão exclusivamente linear, o que torna o sistema mais flexível;
- RGAs são inerentemente dinâmicas e representam uma ferramenta promissora para tarefas envolvendo modelagem dinâmica, memória e comportamento adaptativo.

3.9. Discussão

O modelo conexionista proposto aqui mostrou ser uma maneira interessante de estudar a dinâmica e a estrutura das redes gênicas e seu papel no processamento de informação da célula. As conclusões apresentadas deixam claro o diferencial em termos de perspectiva e de potencial da nossa proposta em relação a outras abordagens *in silico*. A análise do comportamento dinâmico das variáveis do sistema mostrou como a resposta a um estímulo é distribuída entre os nós da rede e como essa característica emerge de processos evolutivos envolvendo mutações estruturais. Esses resultados sugerem uma explicação para propriedades bem conhecidas de redes gênicas reais, como robustez à

variação paramétrica e funcionamento persistente sob falhas de intensidade moderada (BARKAY & LEIBLER, 1997; ALBERT *et al*., 2000).

Embora tenha sido considerado um modelo com apenas mutações estruturais, não se pretende sugerir que otimização dos parâmetros cinéticos das reações não ocorra *in vivo* ou que ela não é realmente relevante para a evolução de organismos reais. Mas como otimização de comportamento baseado no ajuste fino de um conjunto parcimonioso de reações tende a produzir estruturas mais vulneráveis, otimização baseada em estrutura pode ter sido privilegiada pela seleção natural. Além disso, as constantes cinéticas das reações em organismos são determinadas de maneira discreta, pela seqüência de aminoácidos das proteínas, e as possibilidades de ajuste fino em seus parâmetros são limitadas. A mudança em um único resíduo de um motivo enzimático conservado, por exemplo, vai invariavelmente alterar as constantes cinéticas da reação enzimática de maneira drástica. Nesse caso, uma pequena variação nos resíduos não corresponde a uma pequena variação na performance da reação, e o ajuste fino se torna impraticável.

Através dos experimentos com as bactérias artificiais evoluídas, foi possível verificar a capacidade das redes em exibir propriedades interativas complexas, como persistência e tomada de decisão, que são facetas de um comportamento autônomo. Com efeito, células reais apresentam comportamento autônomo, e os exemplos são numerosos na natureza. Considere o caso de um macrófago, o qual não vaga sem objetivo, mas deliberadamente persegue e devora sua presa[6], ou as células de um embrião, como as *neural crest cells*, que se deslocam de um lugar do organismo em formação para uma outra parte para assumir seu papel específico no desenvolvimento (RENSBERGER, 1996). O comportamento complexo das bactérias virtuais resulta da não-linearidade inerente ao modelo e de sua intrincada estrutura realimentada.

Através da análise das estruturas evoluídas, foi possível ter uma idéia da diversidade de possibilidades de solução para um mesmo problema. Em cada rede, as reações bioquímicas assumem papéis diferentes, compondo estruturas alternativas, mas que são capazes de produzir o mesmo comportamento qualitativo. Observe que aqui, diferente das

[6] O macrófago pode se deslocar intencionalmente utilizando seus pseudópodes. Experimentos mostram que os macrófagos são capazes de detectar a presença de uma bactéria no ambiente através de substâncias químicas que a bactéria emite. Quando isto acontece, o macrófago se desloca em direção à bactéria e a persegue, mesmo que ela esteja se movimentando também (RENSBERGER, 1996).

estruturas biológicas conhecidas, as reações e os nós da rede não têm uma identidade própria. Isto é, quando se estuda um sistema biológico real, cada reação, cada "nó da rede" é um componente único, caracterizado por sua origem, propriedades físico-químicas, função, etc. Como conseqüência, essa noção convencional acaba por limitar a possibilidade de abstrair a verdadeira estrutura do sistema de processamento de informação, onde um tipo de molécula é considerado uma variável como muitas outras, e cujas propriedades físico-químicas são interessantes apenas a partir do momento em que elas determinam a maneira como aquela variável vai interagir com as outras variáveis do sistema. Em outras palavras, uma determinada enzima, por exemplo, não deve ser enxergada apenas como uma molécula *E* capaz de catalisar as reações *X* e *Y* e produzir os compostos *A* e *B* – porque analisá-la nesses termos está relacionado à instância apenas e não ao princípio de funcionamento do sistema –, mas como uma variável que afeta e é afetada por outras variáveis através de determinadas relações pré-estabelecidas, e que faz o papel de intermediar uma resposta, reagindo a um dado fluxo de informação quantitativa: amplificando esse fluxo, suprimindo o estímulo, transmitindo esse fluxo para outros nós, etc. Para isto importa a localização deste nó (numa estrutura em rede) em relação aos mecanismos de sensoriamento e de atuação (isto é, se ele participa diretamente do intermédio de uma resposta ou não), a sua conectividade (que vai se relacionar com a amplitude de sua influência no sistema) e a sua velocidade de resposta (que está ligada, além desses fatores, à força das conexões). Portanto, passa-se de descrições e caracterizações específicas da instância sendo investigada para descrições das propriedades das variáveis, em termos de potencial de resposta a um fluxo de informação, e dessa forma a uma caracterização das variáveis relativa ao seu tipo de papel no funcionamento do sistema como um todo. Um bom exemplo dessa caracterização são os nós chamados *hubs* (nós de alto grau de conectividade em uma rede), uma vez que se sabe que um *hub* está envolvido no controle da dispersão de informação para a rede como um todo, muitas vezes resultando em uma transição de fase (GOLDENFELD, 1992). Veja que, neste caso, a caracterização do nó vem das propriedades relacionadas ao seu potencial de reação a um fluxo de informação, e que não é específico de uma instância, mas genérico. Obviamente, não está sendo proposto que uma nova caracterização das inúmeras variáveis de um sistema biológico vai explicar como ele

funciona, mas que uma mudança de foco na abordagem é necessária. A perspectiva de rede permite essa nova abordagem e isso ficou claro pelos experimentos realizados.

A diversidade de estruturas obtidas possibilitou perceber também que redes maiores tendem a ser mais robustas, dado que suas respostas são em geral mais distribuídas, porque no crescimento das redes novos nós e conexões são adicionados para otimizar a resposta do sistema. Pode-se sugerir então que, para cada nova conexão realmente funcional do sistema, muitas outras conexões e nós devem existir de forma a tornar sua funcionalidade eficiente. Por conseguinte, isso acarretaria num aumento inaceitável do número de nós da rede em função da complexidade da tarefa realizada e do número de variáveis externas e de atuação envolvidas, gerando por fim um corpo de variáveis sub-utilizadas que atuam apenas como mecanismos de ajuste. Entretanto, conjeturo que, em vez de aumentar a sua estrutura de modo a acomodar novas funcionalidades, o sistema utilizaria as suas estruturas já existentes para implementar novas funções à medida que estas forem sendo requeridas, criando relações novas entre os nós da rede que já existem e que, possivelmente, estão subutilizados. O que resultaria daí é que todos os nós tenderiam a ser bastante utilizados, dado que aqueles já saturados não podem ser responsáveis por novas funções e aqueles subutilizados teriam o potencial para participar de novas relações com outras moléculas. Saturação aqui está relacionada a diversos fatores, como capacidade física da molécula de acomodar novas interações sem perder outras completamente, e capacidade de produção/disponibilidade da molécula, isto é, sua concentração dentro da célula. Dessa forma, não haveria conexões funcionais sendo otimizadas pela adição de nós responsáveis apenas pelo ajuste, mas funções sendo realizadas parcialmente por uma variedade de nós, e os nós, por sua vez, sendo responsáveis por uma variedade de funções simultaneamente. Para investigar essa questão, é possível realizar experimentos em que a complexidade da tarefa imposta ao sistema e o número de variáveis externas envolvidas aumentam com o tempo, e verificar como se dá a sua adaptação.

É importante salientar que as análises realizadas e os resultados obtidos aqui só foram possíveis devido à simplicidade da metodologia empregada e às propriedades inovadoras do modelo conexionista proposto. As análises evidenciaram a importância da relação entre dinâmica e estrutura, e como essas duas facetas devem ser analisadas em conjunto para abordar o funcionamento das redes gênicas. O comportamento do sistema e a

sua capacidade de executar tarefas também se mostraram fundamentais, não só para a constituição e organização do sistema em si, mas como meio e referência para analisar o impacto de alterações forçadas em sua estrutura. Neste caso, o sistema virtual apresenta a vantagem de ser passível de manipulação arbitrária e de apresentar magnitude e possibilidades de comportamento limitadas. Por fim, a análise da diversidade de estruturas obtidas deixou claro o papel das interações proteína-proteína no processamento de informação celular. Muitas vezes apenas um gene é suficiente para resolver o problema e as proteínas são responsáveis por quase toda a computação. A emergência de uma estrutura sem qualquer interação reguladora deixa isso mais evidente, à medida que mostra um comportamento sendo regulado puramente por interações protéicas. Assim, pode-se questionar se é realmente plausível biologicamente considerar apenas interações gene-proteína nos processos de regulação, e até que ponto essa simplificação resultaria numa abstração suficientemente razoável para reproduzir as propriedades de interesse das redes gênicas.

Capítulo 4

Osciladores Biológicos e Processamento de Informação

Resumo – A proposta principal deste capítulo é a concepção teórica de um sistema de processamento de informação a partir de um conjunto integrado, coerente e coordenado de osciladores biológicos. Processamento de informação, por sua vez, é visto como a capacidade do sistema em perceber estímulos do ambiente em uma conotação temporal e coordenar respostas coerentes a esses estímulos. Essa capacidade de coordenação no tempo é atingida pela interação dos múltiplos osciladores de acordo com uma estrutura organizada, e pressupõe a existência de interfaces que convertem estímulos quantitativos absolutos em informação freqüencial. Neste capítulo, esta hipótese é elaborada partindo de conhecimentos existentes em neurociência, sinergética e dinâmica de coordenação. Embora seja feita a suposição de que este princípio se aplica aos sistemas vivos em geral, o foco principal é dado às redes gênicas.

4.1. Introdução

Vimos nos capítulos anteriores que o sistema regulador de uma célula é responsável por determinar, em conjunto com influências externas, as variações nas concentrações de suas proteínas. No modelo abstrato proposto no Capítulo 3, foi possível analisar como as reações bioquímicas que ocorrem numa célula implementam equações não-lineares, e que essas equações têm de fato o potencial para realizar mapeamentos entrada-saída e operações dinâmicas até certo ponto complexas.

Sob a perspectiva apresentada, o princípio de funcionamento de uma rede gênica se aproxima bastante da noção clássica de uma rede neural, na qual a informação codificada pelo neurônio é representada pela taxa média de disparo de pulsos elétricos (ADRIAN, 1926), gerando assim equações não-lineares em uma estrutura realimentada. No entanto, contradizendo essa visão clássica, foi provado que a codificação na forma de taxa média de disparo é incapaz de explicar inúmeros fenômenos observados no cérebro, por ser limitada

em termos de flexibilidade dinâmica (MACKAY & MCCULLOCH, 1952) (veja MAASS & BISHOP (1999) para uma revisão de uma série de exemplos reais). Teorias mais recentes tentam contornar esse problema, propondo que a codificação de informação neural poderia assumir outras formas, como sincronização (BRUGGE & MERZENICH, 1973; DE CHARMS & MERZENICH, 1996; RIEHLE *et al*., 1997; VAADIA *et al*., 1995) e relações temporais entre eventos sincronizados (BRAGIN *et al*., 1995; ENGEL *et al*., 1991; GRAY AND SINGER, 1989; PRECHTL *et al*., 1997; NEUENSCHWANDER *et al*., 1996), relações de fase entre as freqüências de disparo (BULLOCK *et al*., 1990; O'KEEFE & BURGESS, 1996, SKAGGS *et al*., 1996) e a variabilidade dos intervalos entre os pulsos (SOFTKY & KOCH, 1993).

Embora nada nesse sentido tenha sido afirmado sobre as redes gênicas, é possível esperar que, assim como no cérebro, a codificação da informação genética e protéica como sendo a taxa média das variações seja insuficiente para explicar a complexidade do processamento de informação da unidade básica da vida, a célula. Sendo assim, uma nova ótica, que vai além da concatenação de operações não-lineares numa estrutura realimentada (abordagem empregada no Capítulo 3), deve ser necessária para compreender o processo de regulação celular e, como será visto adiante, os estudos mais modernos em sistemas complexos e neurociência podem apresentar algumas pistas nesse sentido.

A partir dessa motivação, neste capítulo iremos além da perspectiva apresentada nos capítulos anteriores, tentando criar uma imagem mais ampla do que seria o processamento de informação num sistema vivo, associando-o ao conceito de coordenação, e incorporando aspectos da interface informacional no processo de interação com o ambiente. Destaco que as idéias apresentadas aqui são discutidas no plano conceitual apenas, ou seja, em contraste com os Capítulos 2 e 3, não há implementação de experimentos computacionais neste capítulo, embora sejam apresentadas algumas hipóteses que poderão ser verificadas futuramente. O objetivo principal é propor uma discussão a respeito de uma nova ótica sobre o processamento de informação celular.

Um ponto fundamental considerado aqui é que o princípio básico do processamento de informação é o mesmo para todos os sistemas vivos, independente do substrato em que este princípio é implementado. A teoria de sistemas complexos apóia essa hipótese, sugerindo que os mesmos princípios regem a auto-organização e a complexidade dos sistemas auto-organizados em todos os níveis (BAK, 1997; HOLLAND, 1998). Lembre-se,

por exemplo, da Seção 1.5, que discute o fato de que várias instâncias de sistemas vivos, como as redes gênicas, redes neurais e ecossistemas, apresentam em sua estrutura em rede o mesmo padrão organizacional, do tipo hierárquico modular.

As idéias discutidas neste capítulo são amplamente baseadas na teoria da sinergética (HAKEN, 1983), uma vertente da linha de sistemas complexos que busca explicar a formação e a auto-organização de padrões e estrutura em sistemas em não-equilíbrio. Segundo a sinergética, a auto-organização, descrita como a formação de padrões espaço-temporais em sistemas em não-equilíbrio, é basicamente resultado de um processo de coordenação no tempo. Em outras palavras, a auto-organização surgiria da capacidade de um sistema em entrar em sincronia com o ambiente e, sob essa perspectiva, um sistema vivo existiria como tal pela sua capacidade de estar sincronizado com a informação ambiental e com outros organismos (o que também constitui ambiente).

Generalizando esse conceito, a troca de informação, em última instância, só existe entre os sistemas na forma de padrões temporais, e para que haja comunicação (interação coerente) é necessário que esses sistemas estejam em sincronia. Nesse caso, a transferência de informação seria essencialmente freqüencial, e os padrões espaciais apenas a maneira em que essa informação é codificada. Generalizando novamente, a informação no tempo seria a linguagem de comunicação entre os sistemas e de processamento de informação, e o artifício espacial (a organização no espaço), a instância, a maneira pela qual cada sistema implementa essa linguagem. Sendo assim, sistemas não sincronizados seriam incapazes de trocar informação entre si, e, portanto, impossibilitados de reconhecer a existência um do outro.

A motivação dessa teoria é que, através de um formalismo matemático relativamente simples, é possível unificar uma série de conceitos antes vistos (ou pelo menos tratados como) independentes, como percepção, intenção (deliberação) e aprendizado (KELSO, 1995).

Seguindo, portanto, a idéia básica de que coordenação significa sincronia em todos os níveis de auto-organização (KELSO, 1995), será desenvolvida aqui a concepção de um sistema vivo como um conjunto de osciladores em diferentes modos de sincronia, sendo as variações dessas relações de sincronia o princípio em que se baseia o processamento de

informação. O personagem principal aqui continua sendo a célula e suas redes gênicas, mas parte da argumentação é baseada em preceitos e dados empíricos da neurociência.

Na Seção 4.2 deste capítulo, é apresentada a noção de osciladores naturais. Os osciladores são as unidades básicas da codificação da informação freqüencial considerada aqui, e é interessante observar como eles são lugar comum na natureza. A Seção 4.3 discute a idéia de acoplamento entre osciladores e como esse acoplamento pode eventualmente gerar coordenação. Na Seção 4.4, busca-se mostrar como se dá a interação sistema/ambiente, e o que significa a sincronia segundo essa interação. A Seção 4.5 propõe uma explicação para o que constitui especificamente o processamento de informação em sistemas vivos, e a Seção **Erro! A origem da referência não foi encontrada.** traz alguns comentários gerais como considerações finais.

4.2. Osciladores na Natureza

Toda a teoria proposta aqui pressupõe a existência de osciladores na natureza. Esses osciladores, por sua vez, são eventualmente capazes de entrar em sincronia ou não. Quando se pensa em osciladores naturais, talvez a associação mais imediata que venha à mente seja o átomo. Os átomos são osciladores cujo período de oscilação é determinado pela órbita de seus elétrons. Átomos podem entrar em sincronia, o que, segundo a hipótese considerada aqui, caracterizaria o potencial para transferência de informação entre eles. Mas e em sistemas vivos? Será que os osciladores são realmente suficientemente comuns a ponto de comporem as unidades básicas no processamento de informação em um ser vivo? Veja que o objetivo aqui não é propor como os osciladores surgem na natureza, mas sim, partindo do princípio de que eles existem, encontrá-los em operação.

A. Estrutura básica dos osciladores biológicos

Os dois componentes essenciais para que uma oscilação seja produzida são *i*) um efeito inibitório, que envolve uma ou mais variáveis oscilatórias, e *ii*) uma fonte de atraso nesse circuito de realimentação (FRIESEN & BLOCK, 1984). A Figura 4.1 representa uma ilustração esquemática de um mecanismo oscilador, sendo que há uma dinâmica que regula a evolução no tempo da variável *V*. Esse mecanismo pressupõe a existência de um sinal

excitatório que ativa o sistema, uma dinâmica de equilíbrio assintótico e um elemento atrasador capaz de criar as condições para uma oscilação permanente. Esses três elementos são suficientes para que um sistema possa exibir comportamento oscilatório.

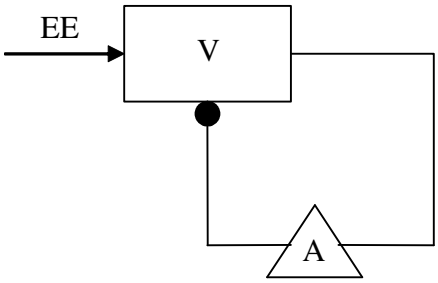


Figura 4.1 Diagrama do mecanismo oscilador. *EE* significa entrada excitatória. *A* representa atraso e *V*, variável. Setas indicam estímulo e círculos indicam inibição.

A idéia básica por trás desse esquema é que a variável *V* inibe o seu próprio crescimento, num efeito de realimentação negativa. Como existe um atraso nessa regulação, o ciclo de realimentação negativa pode gerar oscilação permanente. Suponha que a entrada excitatória é aplicada constantemente e que, inicialmente, o valor da variável aumenta. Chega então um ponto em que a variável atinge um valor tal que faria seu crescimento estabilizar, mas como há atraso na sua regulação, ela continua crescendo (é um processo de inércia). Há, porém, um momento em que o atraso é vencido e a variável começa a decrescer. Quando a variável atinge um ponto em que seu decréscimo poderia cessar, por causa do atraso, ela ainda continua a decrescer (inércia novamente). O atraso é finalmente vencido e o valor da variável começa a aumentar outra vez, recomeçando o ciclo. A variável, então, oscila em torno do que seria o estado de equilíbrio do sistema, caso o atraso não existisse.

B. Oscilador genético

Esse mecanismo de oscilação pode ser facilmente implementado através de um gene cuja proteína regula a sua própria produção, como mostrado na Figura 4.2. A regulação do gene varia não-linearmente com a concentração da proteína reguladora (a regulação segue a função *Hill curve*, como visto no Capítulo 2) e deve haver um atraso entre a produção de proteína e a regulação gênica.

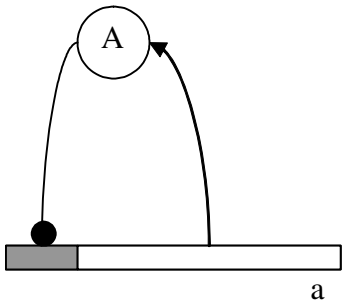


Figura 4.2 Mecanismo oscilador genético simples. A barra representa o gene *a*, que produz a proteína *A*. A parte cinza da barra representa o promotor do gene, que é regulado pela proteína *A*. Setas representam ligações excitatórias e traços com círculos em preto na ponta, ligações inibitórias.

Outro tipo de oscilador genético bastante conhecido são os relógios circadianos (TAKAHASHI & ZATZ, 1982). Um relógio circadiano consiste num mecanismo de sincronização do organismo com o período solar, e pode ser encontrado em praticamente todos os organismos, uni e pluricelulares. O período de oscilação desses relógios, portanto, é naturalmente de 24 horas. O circuito genético mostrado na Figura 4.3 ilustra o princípio básico de funcionamento de alguns dos relógios circadianos bastante conhecidos. Esse mecanismo, obviamente, é genérico, podendo ser implementado em outras estruturas que não genéticas, e é também referido na literatura como oscilador de dois componentes ou oscilador de histerese.

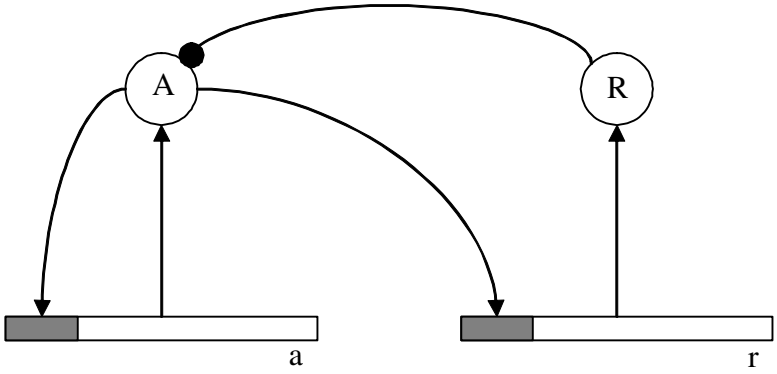


Figura 4.3 Oscilador genético de dois componentes. Letras maiúsculas representam proteínas e letras minúsculas, genes. *A* é a proteína osciladora e *R* a proteína que regula *A*. Setas representam ligações excitatórias e traços com círculos em preto na ponta, ligações inibitórias.

Nesse sistema, a proteína *A* estimula a sua própria produção (realimentação positiva) e também estimula a produção da proteína reguladora *R*, que, por sua vez, inibe a produção de *A*. O efeito da ativação de *A* no próprio gene *a* é mais rápido que a ativação que *A* exerce para a produção de *R* (mais uma vez equações não-lineares estão envolvidas) e, com isso, o sistema exibe oscilação. O gráfico da Figura 4.4 mostra o comportamento das variáveis *A* e *R* no plano de fase do sistema. Veja que a trajetória do sistema converge para um atrator do tipo ciclo-limite (VILAR *et al*., 2002).

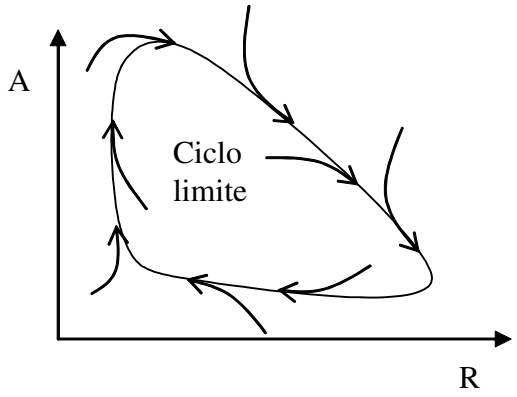


Figura 4.4 Ilustração da trajetória das variáveis protéicas *A* e *R* no seu espaço de fase. Veja que os vários vetores de trajetória convergem para o ciclo-limite.

C. Oscilador glicolítico

Outro tipo de oscilador bastante estudado é o oscilador glicolítico (HESS, 1979), que está relacionado à via metabólica de degradação da glicose. Esse sistema exibe uma periodicidade de aproximadamente 20 minutos e é considerado muito interessante por manter esse comportamento mesmo em ambientes *in vitro*. Não serão apresentados detalhes sobre o funcionamento do sistema, mas observe na Figura 4.5 que a sua estrutura é muito mais complexa e possui muitos circuitos de realimentação.

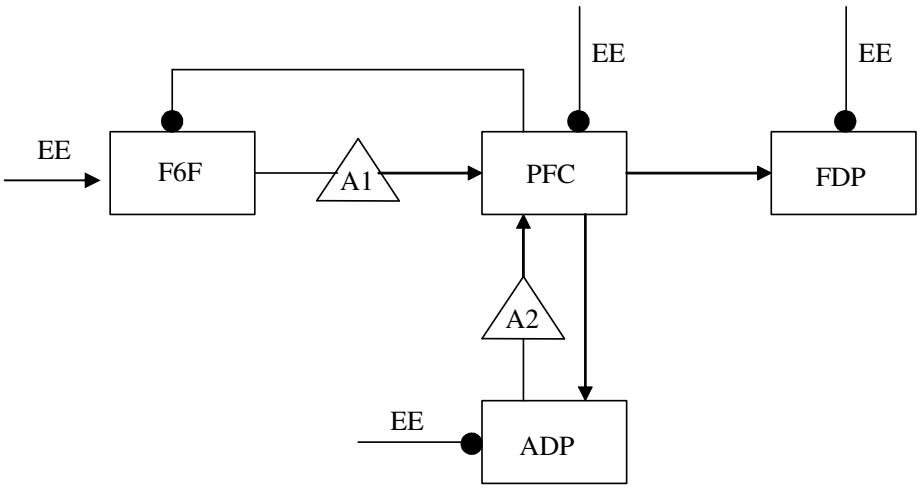


Figura 4.5 Diagrama do oscilador glicolítico. A entrada excitatória é representada pela glicose. As variáveis são F6F (Frutose 6-fosfato), PFC (fosfofrutocinase), FDP (frutose 1,6-difosfato) e ADP (adenosina difosfato).

D. Oscilador neural

Provavelmente o oscilador biológico mais estudado de todos é o oscilador neural (MEECH, 1979). Muitos neurônios têm uma capacidade inerente de disparar pulsos elétricos mesmo sem estímulo externo, mantendo uma freqüência natural constante. A Figura 4.6 mostra o diagrama estrutural do oscilador neural, o qual é regido principalmente pela abertura e fechamento de canais de íons de sódio, cálcio e potássio.

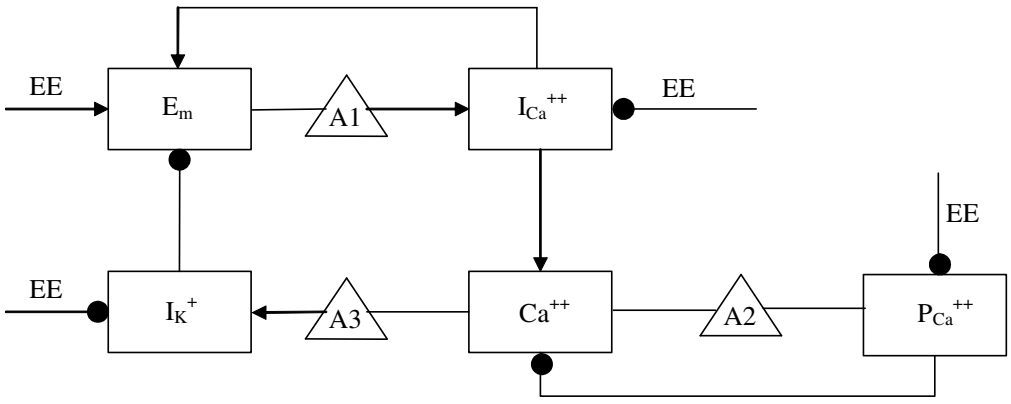


Figura 4.6 Diagrama do oscilador neural. As entradas excitatórias representam as tendências naturais das variáveis. As variáveis oscilatórias são E_m (potencial da membrana), I_{Ca}^{++} (corrente de cálcio através da membrana), I_K^+ (corrente de potássio), Ca^{++} (concentração de íons de cálcio) e P_{Ca}^{++} (atividade da bomba de cálcio).

Um mecanismo muito semelhante é utilizado pelas células cardíacas, as quais também possuem uma freqüência natural de disparo, mesmo na ausência de estímulo, e podem sincronizar para ativar simultaneamente a contração do músculo.

E. Outros osciladores

Além dos osciladores citados acima, há muitos outros na natureza que já foram caracterizados em nível molecular. Veja GOLDBETER (1996) para uma descrição detalhada de alguns deles. Vimos que o sistema nervoso é constituído basicamente de unidades osciladoras, e que também dentro da célula vários mecanismos diferentes, envolvendo genes ou não, podem atuar como osciladores. A hipótese sugerida é que, dentro da célula, assim como no sistema nervoso, a coordenação é regida por um vasto número de componentes osciladores. Obviamente, como esses osciladores podem assumir estruturas completamente diversas, e também complexidades variadas, se torna muito difícil detectá-los, mapeá-los ou isolá-los *in vitro*. Como suporte a essa idéia, é possível citar o trabalho de FRANÇOIS & HAKIM (2004), onde mais de 10 possíveis estruturas ainda inéditas de osciladores genéticos e protéicos são propostas. Essas estruturas foram evoluídas artificialmente em computador.

4.3. Coordenação entre Osciladores

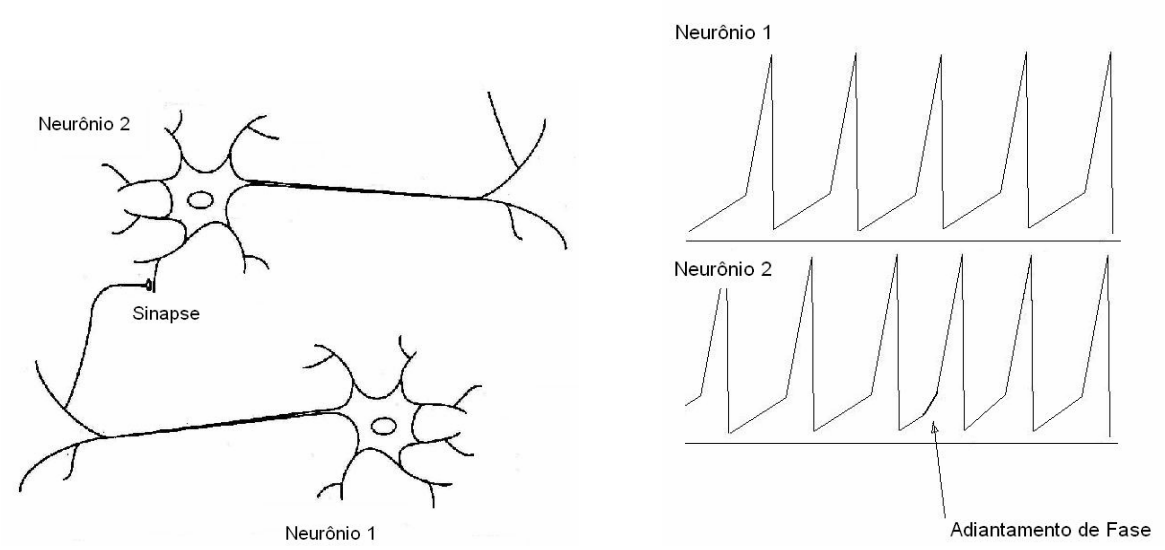
Foi dito que em sistemas naturais há mecanismos que atuam de forma a gerar comportamentos oscilatórios. Mas como esses osciladores naturais podem realizar um trabalho cooperativo, isto é, como eles podem entrar em sincronia? Mais ainda, dado que as células, como os neurônios, por exemplo, nunca são exatamente iguais e que cada uma delas pode possuir uma freqüência natural própria que difere das outras, é possível que elas cooperem e pulsem em fase?

O trabalho cooperativo entre osciladores, mesmo que assimétricos, é possível devido a fatores não-lineares de acoplamento. Será apresentada uma descrição qualitativa desses fatores e suas implicações em termos de comportamento desses sistemas. Análises matemáticas do acoplamento de osciladores envolvendo casos simples podem ser

encontradas em abundância na literatura (GARCIA-OJALVO *et al*., 2004; GONZE *et al*., 2005; KELSO, 1995; LI *et al*., 2006;).

A. Acoplamento entre neurônios

Considere a Figura 4.7(a), que ilustra dois neurônios conectados através de uma ligação sináptica que vai do neurônio 1 ao neurônio 2.



(a) (b)

Figura 4.7 (a) Acoplamento dos neurônios por meio de uma sinapse. (b) Forma de onda do potencial de cada neurônio.

O acoplamento entre eles é determinado pela capacidade de um neurônio em influenciar ou modificar o estado do outro em termos de fase através das sinapses (CARPENTER, 1996). Embora a formulação matemática que caracteriza este acoplamento não seja apresentada aqui, é possível fornecer uma descrição breve e suficientemente intuitiva da dinâmica do processo de interação dos neurônios, que pode levar à sincronia de fase entre eles.

O neurônio funciona como uma bateria capaz de gerar uma diferença de potencial entre o seu meio interno e o meio externo. Esse potencial tende a aumentar até atingir um determinado limiar. Nesse momento, ocorre um pulso de corrente elétrica que é transmitido através de seu axônio para outros neurônios por meio de suas ligações sinápticas e, em seguida, o potencial cai de volta ao seu valor inicial (veja a curva da Figura 4.7(b)). Suponha agora que os neurônios estejam inicialmente defasados, como mostrado no

diagrama de fases da figura. O neurônio 1 dispara primeiro e quando ele o faz, o pulso elétrico que ele emite é transmitido através das sinapses, fazendo com que ocorra uma elevação do potencial do neurônio 2. O resultado disso é que o potencial do neurônio 2 se aproxima mais rapidamente do limiar, fazendo ele disparar mais rápido, e adiantando a sua fase. Isto é, o disparo do neurônio 1 reduz a diferença de fase entre ele e o neurônio 2. Após alguns disparos, os neurônios entram finalmente em fase (como mostrado na Figura 4.7(b)) tendendo a permanecer assim.

Esse é o tipo de acoplamento mais simples possível entre 2 neurônios. Acoplamentos mais elaborados, utilizando sinapses inibitórias e outros artifícios podem gerar padrões de sincronia muito mais complexos, como antifase e ritmos do tipo 1:2, 1:3. 2:3, etc.

B. Acoplamento por sinalização celular

Algumas células, como as amebas *Dictyostelium discoideum* (um dos organismos mais estudados em biologia do desenvolvimento) (GOLDBETER, 1996), são capazes de entrar em sincronia e realizar comportamentos coordenados em uma população inteira. O suporte a esta sincronização está no fato de que as células utilizam uma molécula sinalizadora que influencia o estado da célula e de suas vizinhas, forçando a coordenação.

A grande maioria das abordagens computacionais que tratam da modelagem da sincronia entre osciladores genéticos se baseia na sincronização de uma população de células (GARCIA-OJALVO *et al*., 2004; GONZE *et al*., 2005; LI *et al*., 2006). O princípio básico nesse esquema é que todas as células possuem exatamente o mesmo mecanismo de oscilação e o período dessa oscilação é influenciado pelo próprio sinalizador. Imagine uma população de células do mesmo tipo, todas elas com um mecanismo oscilador de dois componentes, como o mostrado na Figura 4.3. A proteína *A* é a molécula sinalizadora e ela é capaz de atravessar livremente a membrana da célula e influenciar as células vizinhas no ambiente. Experimentos em modelagem computacional mostraram que, dadas essas condições, todas as células da população irão sincronizar após um certo transitório, pois a fase de uma célula influencia diretamente a fase da outra, e o comportamento cooperativo se torna inevitável.

C. Acoplamento entre osciladores intracelulares

Como dito anteriormente, os modelos computacionais que tratam de sincronia de osciladores genéticos utilizam a abordagem intercelular, na qual o acoplamento existe porque todas as células manipulam exatamente a mesma variável (a proteína sinalizadora que pode transitar entre as membranas). Mas e dentro de uma mesma célula? Como se daria o acoplamento entre osciladores uma vez que não existem osciladores idênticos? Em outras palavras, numa abordagem populacional todos os osciladores literalmente manipulam a mesma variável, mas dentro da célula não existem dois genes que produzem a mesma proteína e, portanto, dois osciladores que manipulam a mesma variável não podem existir, o que transforma o cenário em algo qualitativamente diferente.

Segundo a hipótese considerada aqui, o processamento de informação dentro de uma célula se dá através do acoplamento entre diferentes osciladores genéticos e protéicos. Logo, deve haver uma maneira de acoplar dois osciladores, mesmo que eles manipulem variáveis diferentes.

O mecanismo descrito a seguir apresenta um possível artifício para sincronizar dois osciladores diferentes em uma mesma célula. A Figura 4.8 mostra dois osciladores genéticos, semelhantes aos da Figura 4.3, um manipulando a concentração da proteína *A* e o outro da proteína *B*. Cada uma dessas proteínas é capaz de reagir separadamente com uma molécula *C*, formando os dímeros *AC* e *BC*.

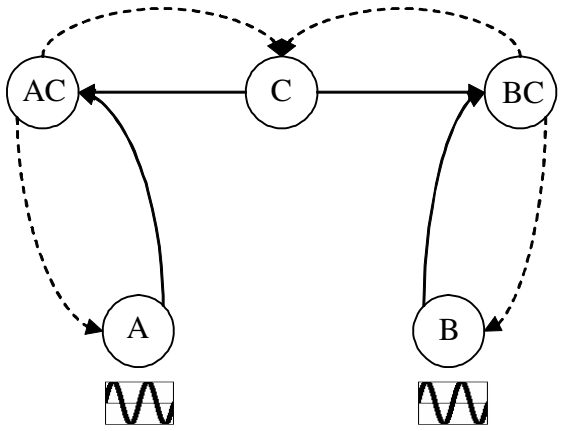


Figura 4.8 Mecanismo de acoplamento entre osciladores genéticos intracelulares. *A* e *B* representam proteínas osciladoras. *C* é uma molécula que reage por dimerização com *A* ou *B*, formando *AC* e *BC*, respectivamente. Os arcos representam o sentido das reações. Reações de ida são representadas por linhas contínuas e reações de volta, por linhas tracejadas.

Para entender o funcionamento do mecanismo, suponha que os osciladores *A* e *B* estão inicialmente fora de fase e que já existe uma certa concentração no ambiente das moléculas *C*, *AC* e *BC*. A concentração de *A* então aumenta devido à sua oscilação natural, gerando mais composto *AC* e reduzindo assim a quantidade de *C*. Como resultado, a reação de formação de *BC* é desequilibrada no sentido de volta. *BC* então começa a se degradar e a aumentar a quantidade de *C* e de *B*. Como conseqüência, a fase de *B* é adiantada, sendo, portanto, atraída para a fase de *A*. Algo semelhante pode ser esperado para o momento em que a concentração de *B* aumenta, sendo a fase de *A* atraída para a fase de *B*. É possível esperar que, após um transitório, os osciladores entrarão em sincronia, e, dependendo dos parâmetros do sistema, esta sincronia será em fase ou anti-fase, e em diferentes razões, como 1:2, 1:3, etc. Obviamente, o comportamento do sistema precisa ainda ser verificado experimentalmente.

Note que o custo de implementação desse mecanismo é baixo, isto é, basta haver uma mesma molécula capaz de dimerizar com outras duas, algo bastante comum dentro duma célula. Essa solução ainda não foi considerada na literatura no contexto de osciladores genéticos.

D. Modelo Haken-Kelso-Bunz

Uma maneira bastante conveniente de visualizar o estado de um sistema de osciladores é através de uma superfície de energia, ou diagrama de potencial. Entretanto, diferente das superfícies de energia tradicionais para sistemas dinâmicos, em que as bacias de atração correspondem a estados de regime das variáveis, aqui os osciladores são vistos em conjunto, e as bacias de atração são estados de sincronia entre eles.

Para isso, ao invés de analisar a fase de cada oscilador individualmente, adotaremos uma *variável coletiva*: a fase relativa entre os osciladores, φ. A Figura 4.9 mostra a superfície de energia gerada pelo modelo didático Haken-Kelso-Bunz (HKB) (KELSO, 1995), no qual dois osciladores simétricos (idênticos) estão acoplados e tendem a entrar em sincronia. As equações do modelo não são mostradas aqui, mas perceba que os osciladores podem coordenar tanto em antifase como em fase, sendo que a sincronia em fase é mais estável por representar um mínimo mais profundo. Portanto, quando exposto a ruído, por

exemplo, o sistema pode passar espontaneamente de uma sincronia do tipo antifase para uma do tipo fase, mas o contrário não é esperado para níveis baixos de ruído.

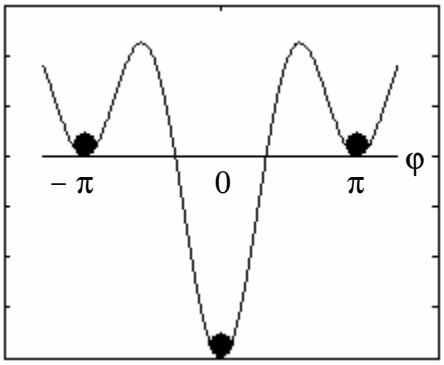


Figura 4.9 Superfície de energia representando os estados de sincronia do sistema. Os círculos negros representam os estados de equilíbrio, onde o sistema apresenta sincronia. O eixo *x* representa a variável φ, isto é, a fase relativa entre os osciladores, e o eixo *y*, o valor da energia.

Agora, veja o que acontece à superfície de energia quando a força de acoplamento entre os osciladores é gradualmente reduzida, para este modelo. A Figura 4.10 ilustra esse processo. Note que os pontos de equilíbrio em antifase vão se tornando cada vez menos estáveis (Figura 4.10(a) e (b)) até se tornarem instáveis (Figura 4.10(c) e (d)).

Cada ponto de equilíbrio do sistema é, na verdade, um atrator do tipo ciclo limite no espaço de estados, mas essa representação facilita a compreensão do fenômeno de sincronização entre osciladores quando uma força de acoplamento existe entre eles. É interessante observar também como os parâmetros do sistema determinam a conformação da superfície de energia e, portanto, como o sistema tende a se comportar ao longo do tempo.

A questão principal aqui é compreender o que significa a sincronia entre os osciladores sob essa ótica e como ela pode ser representada através de uma variável coletiva. Portanto, não importa em que substrato os osciladores estão implementados, mas se existe acoplamento existe interação, e esta interação vai gerar coordenação ou não, dependendo dos parâmetros do sistema. O tipo de coordenação, por sua vez, vai depender das possibilidades existentes, dados esses parâmetros, e do estado inicial do sistema. Todos os osciladores apresentados aqui podem ser analisados sob o mesmo formalismo, o que mostra que a natureza da informação relevante é sempre freqüencial.

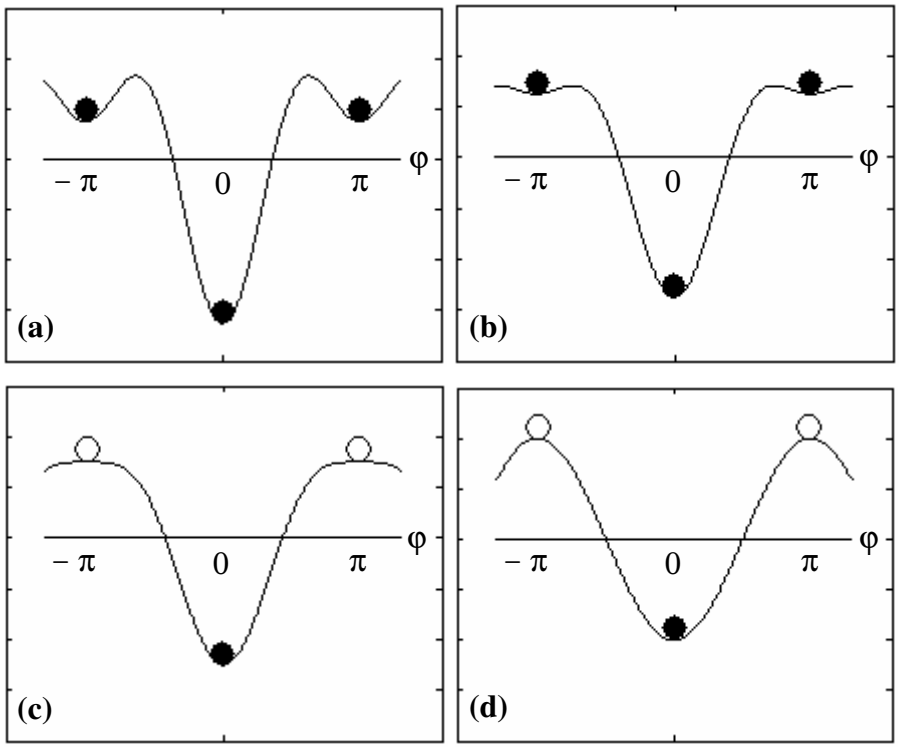


Figura 4.10 Conformação da superfície de energia quando a força de acoplamento entre os ociladores é reduzida gradualmente de (a) a (d). Círculos preenchidos representam pontos de equilíbrio estável e os círculos em branco, pontos de equilíbrio instável.

4.4. Coordenação com o Ambiente

Supondo que o processamento de informação em um organismo vivo se dê como um conjunto de osciladores acoplados de diferentes modos e capazes de, através de suas interações, realizar associações entre estímulos e coordenar reações no tempo, é preciso determinar como esse sistema interage com o ambiente de forma a retirar as informações crucias à sua sobrevivência.

Com efeito, a subsistência de um sistema vivo está associada ao ambiente. É o ambiente que fornece o fluxo de matéria e energia necessárias para a sua integridade e para a manutenção do seu estado de não-equilíbrio. A hipótese considerada aqui (assim como também sugerida em KELSO (1995)) é que, para que a auto-organização em sistemas vivos exista, é preciso que o sistema entre em sincronia com o ambiente. Assim, ele pode adquirir

as informações necessárias para sobreviver. No entanto, a idéia de sincronizar com o ambiente pode ter duas facetas diferentes, e deve ser analisada com cautela.

A. Quando a informação do ambiente é naturalmente freqüencial

A maneira mais imediata de compreender a sincronia com o ambiente é no caso em que a informação proveniente deste é naturalmente freqüencial. Considere, por exemplo, os relógios circadianos. A informação do ambiente, neste caso, é freqüencial, dado o período de 24 horas do dia, e o organismo deve entrar em sincronia com esta informação para garantir a sua subsistência. Quando se trata de um organismo que realiza fotossíntese, por exemplo, estar sincronizado com o período solar pode ser de fato determinante para a sua integridade.

Uma vez que a subsistência do organismo está associada a retirar informações do ambiente e reagir de alguma forma a essas informações, é necessário primeiro que este organismo se ajuste aos padrões temporais do ambiente (que podem ser bastante complexos). Estar em sincronia com esses padrões significa conhecer esses padrões; só assim será possível modular uma resposta coerente a esta informação de entrada.

B. Quando a informação do ambiente não é freqüencial

É possível citar outras situações em que a informação do ambiente é freqüencial (como o caso das ondas sonoras, cujas freqüências mecânicas são convertidas em freqüências de pulsos elétricos (BRUGGE & MERZENICH, 1973)), mas é provável que para sistemas vivos essa circunstância seja, na verdade, uma exceção. As respostas de um organismo devem, sim, ser ponderadas no tempo e coordenadas com outros estímulos e outras respostas, mas esses estímulos e respostas nem sempre possuem caráter freqüencial, embora possuam a sua localização no tempo.

Supondo que o processamento de informação de um organismo vivo é composto de uma série de osciladores acoplados de diferentes maneiras, como é possível então que esse sistema interaja coerentemente com informações de caráter quantitativo, não-freqüencial?

A solução apresentada para isto é simples. Para que haja sincronia neste caso, basta que a informação quantitativa do ambiente seja modulada de alguma forma em informação freqüencial. Ou seja, deve haver algum conversor no sistema que transforma a informação

do ambiente em um sinal freqüencial. Informações mais intensas seriam traduzidas em freqüências mais altas e informações menos intensas em freqüências mais baixas (o contrário também é possível). Para isso, deve haver a possibilidade de modular a freqüência de um oscilador baseado na intensidade da informação de entrada.

Há várias outras maneiras de se modular a freqüência de um oscilador. Considere o oscilador genético de dois componentes da Figura 4.3. Para alterar a sua freqüência de oscilação, basta manipular qualquer uma de suas variáveis, através de, por exemplo, reações que modificam a concentração da proteína *A* ou *R* diretamente, e proteínas reguladoras que se ligam aos genes, alterando a taxa de síntese de proteínas.

Em neurociência, a idéia de modulação de um sinal sensorial em diferentes freqüências no sistema nervoso é bem conhecida e já foi comprovada em vários contextos. Em um experimento pioneiro utilizando sapos, ADRIAN (1926) mostrou que os sensores que monitoram o estiramento da perna desses animais produzem um sinal em freqüência em função do estímulo externo (no caso, um peso que força o estiramento do membro). A freqüência emitida pelos neurônios varia linearmente com o aumento da carga, mas esse efeito é saturado para cargas mais elevadas. Efeitos semelhantes foram encontrados, por exemplo, para estímulos visuais relacionados à intensidade de sinais luminosos (HUBEL & WIESEL, 1962).

A Figura 4.11 ilustra a proposta de interface sistema/ambiente. Nem todos os componentes da figura precisam necessariamente estar presentes, esse é apenas um esquema genérico.

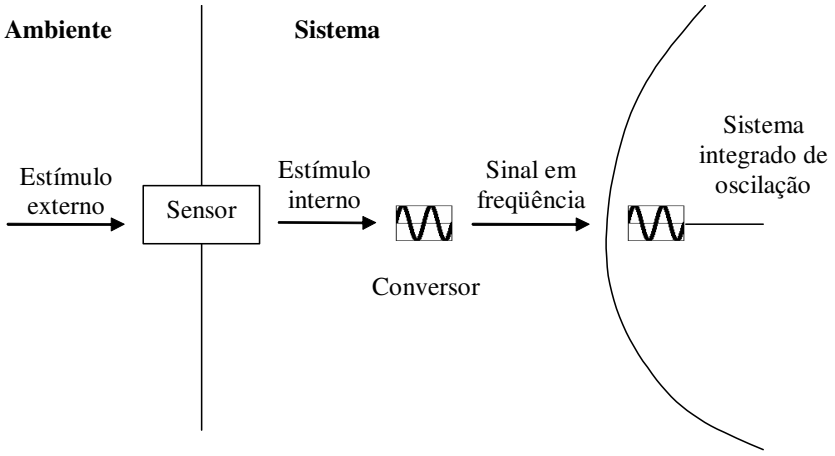


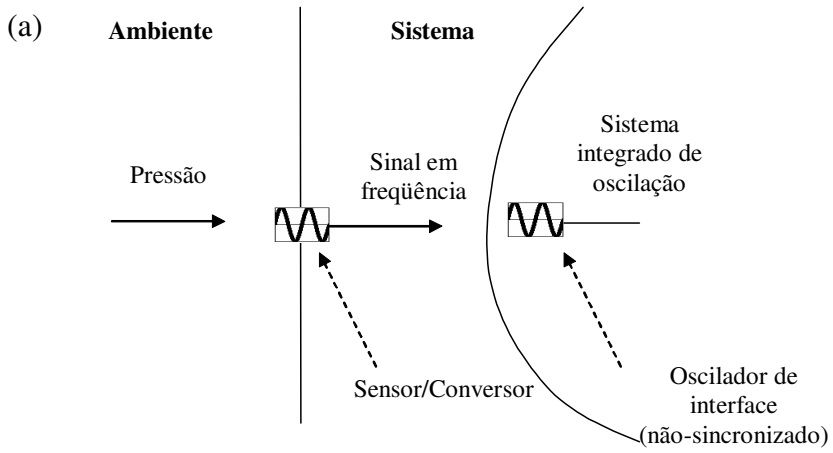
Figura 4.11 Esquema ilustrativo da interface entre o sistema e o ambiente num processo de percepção.

Segundo este esquema, o estímulo externo é captado pelos sensores, que simplesmente repassam a informação ao sistema, mas a intensidade do sinal repassado vai depender das propriedades dos sensores. Esse estímulo interno é então convertido em freqüência através de um oscilador e agora se torna passível de interpretação pelo sistema como um todo, chamado aqui de *sistema integrado de oscilação*. A razão pela qual o conversor é separado do resto do sistema é que, dependendo da freqüência do sinal que ele produz, será possível que o sistema entre em sincronia com este sinal ou não. Entrar em sincronia significa que o sinal influencia de alguma forma a coordenação do sistema como um todo, isto é, há transferência de informação.

C. Caso de estudo 1: tato

Vamos considerar agora um caso de estudo que tenta ilustrar a dinâmica desse processo de coordenação com o ambiente. Será apresentada uma possível concepção da interação de um sistema com a informação ambiental.

A Figura 4.12 mostra o processo de interação quando um organismo sensível ao tato é estimulado através do meio externo. Inicialmente (Figura 4.12(a)) uma pressão física é exercida pelo ambiente sobre o organismo e este sinal é, então, convertido em uma freqüência de oscilação. No entanto, essa freqüência é baixa demais e não é capaz de gerar sincronia entre o sistema integrado de oscilação e os sensores. Não há, portanto, percepção do sinal.



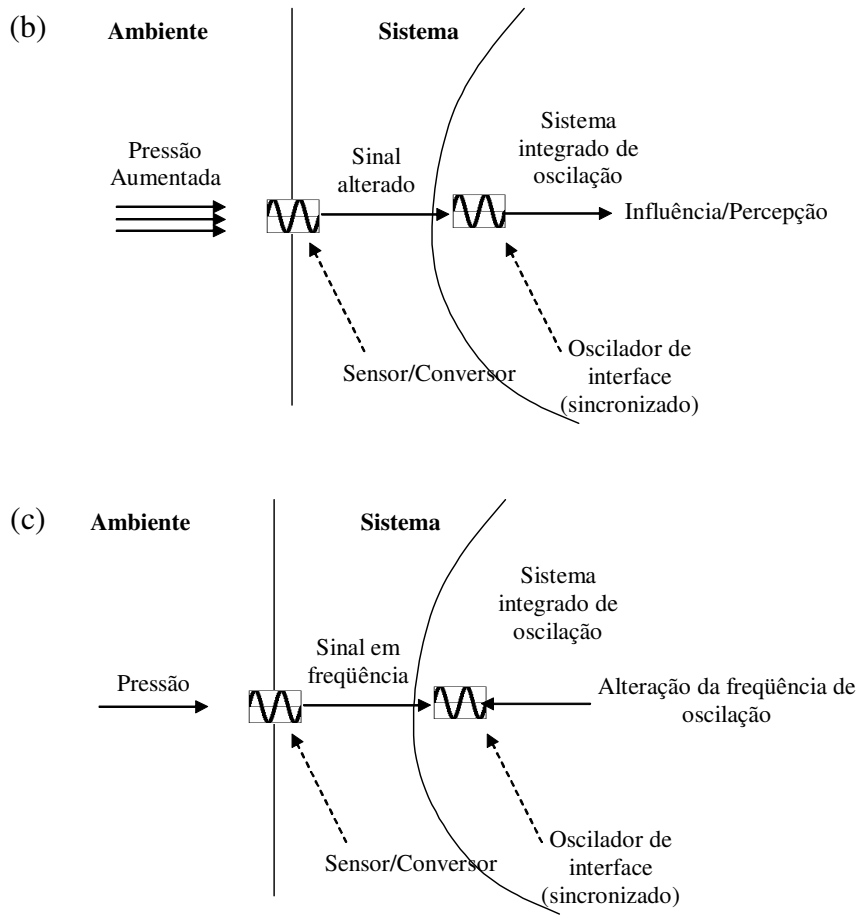


Figura 4.12 Interação sistema/ambiente na percepção do tato.

Numa segunda situação (Figura 4.12(b)), a pressão exercida pelo ambiente é aumentada, fazendo então o sinal interno aumentar de freqüência. Agora, as freqüências do sensor e do oscilador interno são mais compatíveis, e a sincronia é possível. Observe que o fato de estar sincronizado causa uma alteração no sistema, porque a sincronia é, na verdade, um meio-termo entre a freqüência de oscilação natural do oscilador de interface e a do sinal gerado pelo conversor. Uma vez que há acoplamento entre essas estruturas, o estado de equilíbrio do sistema é, em geral, resultado do adiantamento da fase de um oscilador e o atraso da fase do outro. Há, portanto, uma alteração na freqüência original do oscilador de interface, e essa alteração é, naturalmente, propagada para o resto do sistema. A partir deste ponto, modificações na intensidade do sinal de entrada serão repassados em termos de alteração de freqüência para o resto do sistema (há comunicação constante), até que a sincronia não possa mais ser mantida e a coordenação seja perdida e não haja mais

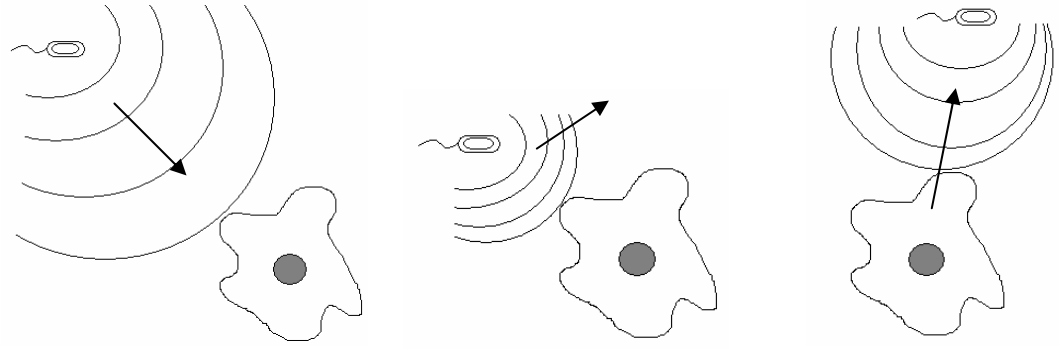
transferência de informação. Esta sincronia entre o sinal e o oscilador interno será chamada, portanto, de percepção.

Numa terceira situação (Figura 4.12(c)), a pressão do ambiente continua a mesma da situação da Figura 4.12(a), mas, devido a um estímulo interno, a freqüência de oscilação do oscilador de interface é alterada e, mesmo com um estímulo fraco do ambiente, a sincronia agora se torna possível. Esse estímulo interno pode ser resultado direto ou indireto de um ou vários estímulos externos, ou, em altíssimo nível, pode ser considerado como resultado da intenção ou deliberação do organismo. Em KELSO (1995) é mostrado de forma consistente que a dinâmica intencional corresponde a uma modificação, por meios internos, da superfície de energia do acoplamento entre osciladores. Alterar a freqüência de um dos osciladores produz exatamente este efeito. Portanto, percepção, segundo esse modelo, pode resultar tanto de um estímulo externo quanto de um estímulo interno.

D. Caso de estudo 2: quimiotaxia

O segundo caso analisado é uma tarefa de quimiotaxia realizada por um macrófago. O macrófago possui sensores capazes de detectar elementos químicos liberados por uma bactéria, e pode se deslocar em direção à bactéria e fagocitá-la guiado por esses estímulos químicos.

A Figura 4.13(a) mostra a situação inicial em que a bactéria está posicionada relativamente longe do macrófago e o estímulo que chega até ele é fraco.



(a) (b) (c)

Figura 4.13 Macrófago interagindo com uma bactéria. Setas indicam vetores de direção. (a) O sinal químico emitido pela bactéria é muito fraco e o macrófago não é capaz de notá-la. (b) A bactéria se aproxima e o sinal químico emitido por ela é suficientemente forte para gerar percepção. (c) O macrófago reage se deslocando em direção à bactéria.

Na Figura 4.13(b), a bactéria está mais próxima do macrófago e o estímulo é suficiente para gerar sincronia com os osciladores internos, o que indica que o macrófago percebe a presença da bactéria. A seguir, na Figura 4.13(c), a bactéria se afasta do macrófago, e o enfraquecimento do sinal tende a reduzir a freqüência do oscilador interno. Essa redução é repassada para o resto do sistema, modificando o seu estado interno e modulando uma resposta em forma de deslocamento do macrófago em direção à bactéria. Se essa resposta for coerente, a sincronia tenderá a ser mantida, pois o deslocamento do macrófago fará o sinal permanecer forte.

Note que a transferência de informação para o sistema é possível por causa da sincronia entre o estímulo em sua forma freqüencial e o oscilador de interface. O acoplamento permite que, mesmo com modificações na freqüência do sinal, ainda seja possível a sincronia, e essas modificações vão resultar em transferência de informação para o restante do sistema.

Um ponto importante é que, à medida que o sinal aumenta de freqüência, ou mesmo diminui, a sincronia não é necessariamente perdida, mas pode ser mantida em uma outra razão de freqüências. Portanto, no caso da aproximação do macrófago, no momento em que a sincronia é estabelecida ela pode ser do tipo 1:5, e à medida que o sinal aumenta a razão pode eventualmente se estabilizar em 1:3. Essa mudança discreta é transferida ao sistema. Ainda assim, à medida que a freqüência aumenta e a razão é mantida, o fato de modificar continuamente a freqüência do oscilador de interface pode resultar em modificações discretas, em termos de razão de sincronia, em outras partes do sistema que estão acopladas indiretamente a este oscilador, caracterizando, portanto, mudanças de estado do sistema. Se a freqüência do sinal atinge uma faixa em que a sincronia não é mais possível, a relação de fases entre o oscilador de interface e o sinal se torna caótica, e não há mais coordenação (KELSO, 1995).

E. Percebendo o mundo

Para resumir como se dá no organismo a percepção do mundo através dos sensores (uma propriedade que é traduzida muito bem pela palavra em inglês *situatedness*) considere a Figura 4.14.

Esse esquema supõe a existência de duas interfaces, uma que separa o ambiente do meio interno, mediada pelos sensores e conversores, e uma que separa as informações do ambiente e o funcionamento do sistema como um todo. Obviamente, essa segunda interface não precisa existir realmente; tudo o que é interno é componente do sistema. A separação é esquematizada aqui apenas para ilustrar o fato de que a informação ambiental pode estar influenciando ou não o sistema através de sincronia. A informação dos sensores pode, inclusive, estar em sincronia com os osciladores de interface, mas pode estar sendo barrada num nível mais interno da cadeia de acoplamentos.

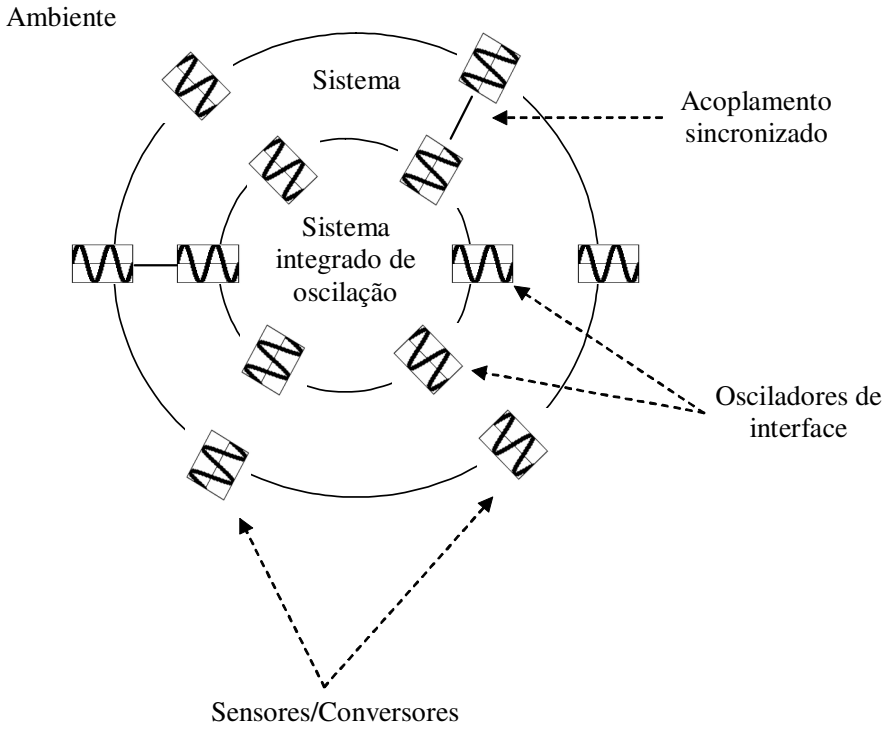


Figura 4.14 Esquema ilustrativo da interface entre sistema e ambiente.

Portanto, em relação à capacidade de percepção do ambiente, a complexidade do sistema está relacionada ao número de variáveis sendo monitoradas, à variabilidade em

freqüência que cada um dos sinais pode assumir, e à capacidade do sistema em se manter em sincronia com esses sinais mesmo que eles variem (se adaptando às suas mudanças), seja passivamente, por meios internos apenas, ou ativamente, através de ações que modifiquem o meio.

4.5. Processamento de Informação

Vimos como é feita a interação do sistema com o ambiente segundo o modelo proposto. Agora vamos olhar mais de perto o processamento de informação em si, e como é constituído o sistema integrado de oscilação.

Um organismo deve ser capaz de realizar tarefas complexas através da cooperação de seus osciladores, e grande parte dessas tarefas deve ocorrer em paralelo. Como as tarefas diferem em sua natureza, cada uma vai ter o seu ritmo particular. Portanto, por mais que consideremos um sistema em sincronia, é preciso haver diversidade nessa sincronia, e a organização do sistema deve permitir isso. Ademais, essas tarefas realizadas por meio de estruturas de baixo nível devem, também, ser passíveis de serem coordenadas em alto nível, gerando assim atividades mais complexas ainda.

Outro ponto a ser considerado é que a flexibilidade dos osciladores é geralmente limitada. Cada oscilador possui uma freqüência natural, mas através do seu acoplamento deve ser possível gerar freqüências completamente novas para garantir flexibilidade ao sistema.

A. Estrutura da coordenação

Considere um oscilador genético do tipo mostrado na Figura 4.3. A dinâmica desse sistema mostra que cada uma de suas variáveis será um oscilador, e não só a proteína *A*. Além disso, qualquer produto de uma reação que envolva uma dessas variáveis também será um oscilador em potencial. Imagine, por exemplo, uma proteína *B* que se liga à proteína *A*. O dímero resultante, *AB*, também vai oscilar junto com a variável *A*. Isso vai alterar a freqüência natural do oscilador, e se esse sistema em conjunto vai gerar um comportamento oscilatório coordenado ou não, vai depender dos seus parâmetros. Agora, considere que uma outra proteína se liga a *AB*, formando *ABC*. Novamente, é possível que

o comportamento gerado seja coordenado, e isso vai depender dos parâmetros, mas a probabilidade de que esse acoplamento em seqüência gere coordenação se torna mais restrita ainda. E esse efeito pode ser generalizado para mais reações acopladas em cadeia.

Daí é possível levantar dois pontos importantes, para o caso das redes gênicas e protéicas:

1) Nem todos os componentes do sistema precisam ser necessariamente osciladores para realizar comportamento periódico, embora os osciladores sejam necessários como força motora;
2) Muitos acoplamentos em seqüência dificilmente vão gerar coordenação. É preciso uma estrutura mais adequada.

Considere agora a seguinte proposta de estrutura (Figura 4.15), onde vários osciladores estão acoplados com razões variadas. Essa estrutura está de acordo com as estruturas modulares mapeadas em redes protéicas (RAVASZ *et al*., 2002). Não há na figura distinção entre o que são realmente osciladores e o que são proteínas ligadas a esses osciladores através de reações bioquímicas.

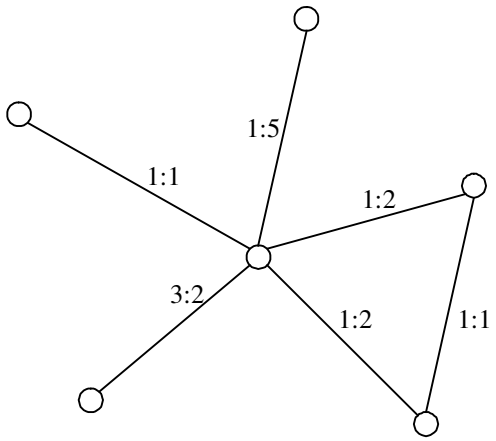


Figura 4.15 Estrutura de uma rede de acoplamentos de osciladores protéicos em sincronia. Cada nó da rede corresponde a uma proteína (variável osciladora). Os números correspondem às razões de freqüência dos acoplamentos.

Essa estrutura radial tende a ser mais estável que osciladores simplesmente acoplados em série, pois a sincronia entre dois osciladores reforça a periodicidade e, portanto, as outras sincronias. Além disso, ela permite diversidade de razões de freqüência, enquanto uma estrutura em seqüência seria muito limitada nesses termos. Embora possa ser

difícil conceber vários osciladores operando em perfeita sincronia, é aceitável que eles estejam trabalhando em coordenação relativa, um conceito que será descrito na próxima subseção e que consiste num regime muito mais flexível. Obviamente, as propriedades sugeridas para esta estrutura, embora aparentemente intuitivas, requerem comprovação experimental.

Repare para o detalhe na figura de que dois osciladores da periferia estão conectados entre si. Esse acoplamento reforça a estabilidade da estrutura, mas note que ele nem sempre vai ser possível, vai depender das razões de freqüência em que eles estão sincronizados com o oscilador central. Como resultado dessa estrutura, todo o conjunto pulsa num mesmo ritmo. Embora cada um dos osciladores possua uma razão de freqüência particular, todos eles oscilam em função do oscilador central. É ele quem dita o ritmo e, ao mesmo tempo, seu ritmo é ditado pelo conjunto de osciladores acoplados a ele.

Agora, considere a Figura 4.16, onde uma rede maior, mas com uma estrutura coerente com a da Figura 4.15, é apresentada.

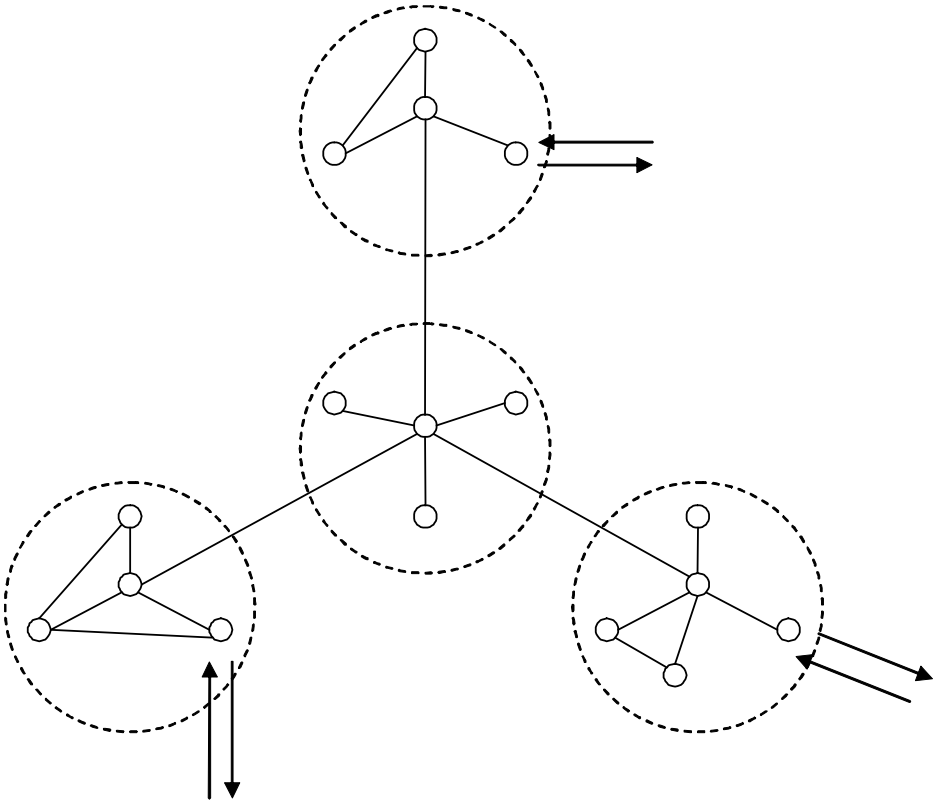


Figura 4.16 Rede de osciladores acoplados. Círculos tracejados destacam módulos funcionais. Setas indicam interação com o ambiente por meio de sensores ou atuadores.

Nessa rede, os círculos em tracejado destacam os módulos funcionais. Eles possuem uma estrutura semelhante à da Figura 4.15. Cada um desses módulos realiza uma operação específica através da coordenação da atuação de proteínas. Cada um deles pulsa num ritmo diferente e a coordenação entre eles é regida pelo oscilador central (e, ao mesmo tempo, a freqüência do oscilador central é regida por eles). Com isso, a coordenação em baixo nível acaba por gerar coordenação em alto nível, pois cada módulo como um todo se tornou um oscilador. Através do oscilador (ou módulo) central esses módulos podem entrar em sincronia entre si ou não, e com variadas razões de freqüência. Assim, é possível coordenar tarefas muito complexas através do mesmo mecanismo que coordena tarefas simples, e a complexidade cresce à medida que a quantidade de níveis hierárquicos[7] cresce também. Em um outro nível, toda essa rede pode ser considerada como um novo módulo oscilador. A estrutura fractal, hierárquica modular, encontrada nos mapas de estrutura de rede de sistemas vivos (BARABÁSI, 2002), cabe perfeitamente aqui.

As setas na figura representam interação com o ambiente através de sensores e atuadores. Podemos agora conectar essa estrutura com a da Figura 4.14. A rede mostrada aqui é o sistema integrado de oscilação e as proteínas associadas a setas são os osciladores de interface. Através da interação com o ambiente, a freqüência desses osciladores tende a se modificar e isso pode resultar numa mudança das razões de sincronia de outros osciladores do módulo ou até dessincronização. Isso, eventualmente, pode alterar a freqüência do módulo inteiro e culminar numa alteração global no sistema. Assim, modificações locais podem afetar todo o sistema.

Portanto, nesse sistema idealizado as relações de sincronia estão se alterando o tempo todo. Módulos podem operar em conjunto ou não, assim como as proteínas, dependendo do instante e em função dos estímulos externos, mesmo que indiretamente. Aliás, quanto mais alto o nível hierárquico, mais indiretas serão as influências externas, e a mudança das relações de sincronia nos níveis mais altos assumirá uma conotação quase que autônoma.

[7] A estrutura é hierárquica no sentido de que um conjunto de elementos em um nível inferior forma um nível superior, e um conjunto desses elementos de nível superior forma um nível mais elevado ainda, e assim sucessivamente. A estrutura é chamada hierárquica modular, pois um módulo é formado por vários módulos menores que, por sua vez, são formados por módulos menores ainda. Devido a esta auto-similaridade, essa estrutura é chamada também de fractal.

Nesse ponto é interessante formalizar dois conceitos que estão sendo tratados aqui:

- *Dinâmica de primeiro nível*: Diz respeito à dinâmica de interação de osciladores acoplados e seus estados de sincronia ou não.
- *Dinâmica de segundo nível*: Refere-se a como as relações de sincronia (dinâmica de primeiro nível) se alteram ao longo do tempo; é uma meta-dinâmica que descreve a evolução do sistema.

A idéia de módulos e sub-módulos em sincronia no cérebro é conhecida na literatura de neurociência (SINGER & GRAY, 1995; ENGEL *et al*., 1997). Há evidências de que a sincronia entre diversos módulos provê uma forma de cooperação na qual várias características são associadas a um mesmo estímulo. Essa noção está de acordo com a proposta apresentada acima.

B. Modulando freqüências

Nesta seção, uma possível solução para gerar freqüências completamente novas a partir de osciladores de flexibilidade limitada é apresentada. Mas antes de ir direto à proposta desse mecanismo, é necessário introduzir o conceito de coordenação relativa.

Coordenação relativa é um fenômeno pouco estudado e é vista como uma solução encontrada pela natureza para realizar sincronia quando os osciladores envolvidos são assimétricos. Considere o seguinte exemplo. Imagine dois adultos andando e conversando ao mesmo tempo. Dado que o tamanho de suas passadas é aproximadamente a mesma, é bem provável que esses adultos caminhem em perfeita sincronia de seus passos. No entanto, quando um adulto caminha enquanto conversa com uma criança, ambos também tenderão a andar em sincronia, mas dado que as passadas diferem bastante, a criança algumas vezes terá de dar dois passos ao invés de um só, de forma a acompanhar o adulto e manter a sincronia. Esse fenômeno caracterizado pela sincronia em grande parte do tempo, e perda da sincronia e sua retomada rápida, é o que chamamos coordenação relativa.

Os diagramas de potencial para um sistema assimétrico segundo o modelo HKB, discutido anteriormente, ilustram esse fenômeno com mais clareza. A Figura 4.17 mostra a curva de potencial do sistema HKB quando a assimetria dos osciladores cresce gradualmente.

Veja que a assimetria distorce a curva de potencial, tornando a sincronia em uma das possíveis antifases mais estável e a outra menos estável (Figura 4.17(a)). Quando a assimetria aumenta, os pontos de equilíbrio em antifase acabam por se tornar instáveis (Figura 4.17(b)). O mesmo acontece para o ponto de equilíbrio em fase, na Figura 4.17(c), e agora os pontos de equilíbrio em antifase desaparecem. Na Figura 4.17(d), não há mais pontos de equilíbrio.

Considere agora uma esfera rolando no eixo φ com velocidade constante sobre a curva de potencial da Figura 4.17(d). Embora não haja mais pontos de equilíbrio, é natural supor que a esfera vai descer rapidamente e atrasar a sua descida quando se aproxima de $\varphi = 0$, pois essa região da curva é relativamente plana. Se esta esfera representa o estado do sistema, este atraso significa que o sistema está em quase-sincronia. A esfera continua a rolar e quando chega ao final da curva retorna ao ponto de início, pois a curva se repete a partir desse ponto.

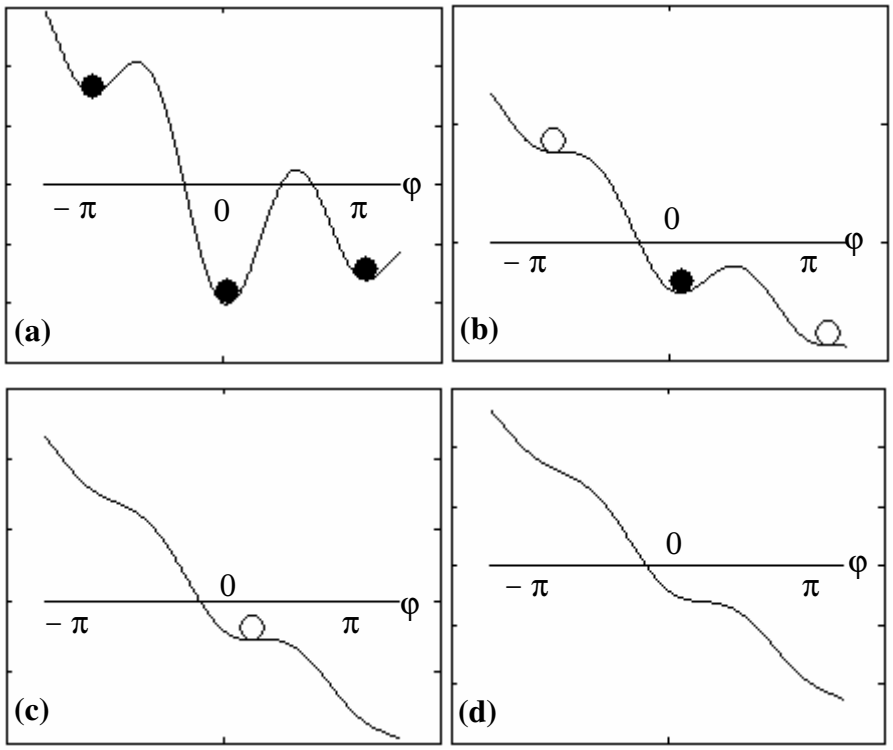


Figura 4.17 Curva de potencial para osciladores assimétricos acoplados. A assimetria dos osciladores aumenta gradualmente de (a) a (d). Círculos preenchidos representam pontos de equilíbrio estável e os círculos em branco, pontos de equilíbrio instável.

A dinâmica do sistema é então caracterizada por momentos duradouros de quase-sincronia e momentos rápidos de dessincronia, em que a relação de fases adianta em π. Esse tipo de efeito, próprio da coordenação relativa, é chamado intermitência. A Figura 4.18, a seguir, dá uma outra ilustração do comportamento intermitente para o mesmo sistema.

Há outras maneiras que podem ser empregadas para ilustrar o comportamento intermitente, mas a descrição fornecida até agora já é suficiente para concebermos nosso mecanismo de produção de novas freqüências.

Mais uma vez, a solução apresentada aqui é bastante simples. Suponha que cada um dos osciladores assimétricos é um oscilador genético, modulando o comportamento das proteínas *A* e *B*, respectivamente. Considere agora uma proteína *C*, produzida por um gene que depende da presença simultânea das proteínas reguladoras *A* e *B* para estar ativo, e que tanto a produção quanto a degradação de *C* têm constantes de tempo grandes em relação a *A* e *B*. Logo, se repararmos na Figura 4.18, a freqüência de *C* (o novo oscilador) corresponderá à freqüência dos platôs de intermitência. Se considerarmos que a freqüência de oscilação máxima e mínima dos osciladores individuais é limitada, temos então a implementação de um oscilador com uma freqüência nova que pode sair dessa faixa. Se a freqüência de *A* e *B* for tal que, durante o platô eles oscilem por períodos completos várias vezes, a freqüência de *C* pode assumir um valor bem mais baixo que as de *A* e *B*. Além disso, a freqüência de *C* pode ser modulada de forma contínua dependendo do estado do sistema de acoplamento entre *A* e *B*, em relação ao ponto de equilíbrio, isto é, de acordo com a largura dos platôs. Essa solução aumenta a flexibilidade do sistema de osciladores e não foi proposta na literatura ainda no contexto de osciladores genéticos.

Outra possível propriedade desse mecanismo é que ele pode ser utilizado para ativar ou desativar módulos funcionais. Suponha que o oscilador central de um módulo funcional é justamente regido pela proteína *C*. Se, por algum motivo, a coordenação relativa entre *A* e *B* for interrompida, a proteína *C* não será mais produzida, e o módulo pode simplesmente perder a sua dinâmica ou o seu funcionamento coerente.

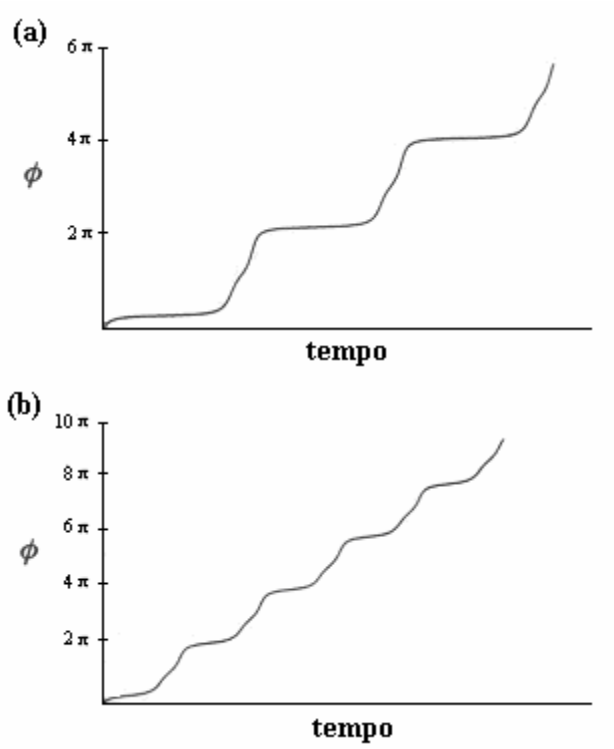


Figura 4.18 Comportamento intermitente da relação de fases para o sistema HKB assimétrico. Os platôs correspondem aos momentos em que os osciladores estão praticamente em fase. Em (a) o sistema está mais próximo a um ponto de equilíbrio estável do que em (b).

4.6. Discussão

A concepção apresentada neste capítulo é inédita no contexto de redes gênicas, mas se mostra bastante coerente com observações na linha de neurociência. Ainda nesta última, a idéia de uma estrutura hierárquica de módulos osciladores é nova, e provê uma explicação funcional não só para os fenômenos de sincronia observados em regiões do cérebro, mas também para como as operações são coordenadas no cérebro como um todo. Obviamente, não é escopo desse trabalho se aprofundar nesse mérito. Portanto, deixemos essa discussão para os especialistas em neurociência e ciências cognitivas.

É interessante perceber também que a estrutura fractal emerge naturalmente do princípio de funcionamento do sistema idealizado acima, e esse é um ponto importante, pois ainda não se sabe qual a relação entre esse tipo específico de estrutura e a dinâmica interna dos sistemas vivos.

A principal dificuldade encontrada na elaboração das idéias propostas foi a incipiência da linha de pesquisa em acoplamento de osciladores biológicos. Quase não há na literatura pesquisas envolvendo acoplamento de múltiplos osciladores assimétricos (LI *et al*., 2006), e não foi possível encontrar nenhum estudo envolvendo razões de acoplamento variadas ou coordenação relativa, ou que explore arquiteturas de rede variadas. As principais referências nessa linha são (ABBOTT & VAN VREESWIJK, 1993; GERSTNER *et al*., 1993; GOLOMB *et al*., 1992; GRANNAN *et al*., 1993; HOPFIELD & HERTZ, 1995; MIROLLO & STROGATZ, 1990; STROGATZ & STEWART, 1993; TERMAN & WANG, 1995; USHER *et al*., 1993). O estudo em acoplamento de osciladores pode ser visto como uma linha de pesquisa bastante promissora e que deve ser considerada como perspectiva futura de investigação.

Capítulo 5

Conclusão

Esta dissertação tratou das redes gênicas e protéicas sob três perspectivas alternativas. No entanto, essas três vertentes de análise são complementares e podem ser consideradas em conjunto no estudo do funcionamento das redes gênicas. A proposta de ferramentas computacionais capazes de inferir estruturas a partir de dados de expressão é fundamental para mapear as interações gênicas e ter acesso às cadeias de relações causais responsáveis pelos fenômenos celulares de interesse. Modelagens computacionais também são requeridas para simular o funcionamento de um sistema regulador sob condições desejadas. Simulações computacionais permitem a manipulação arbitrária do sistema e de suas condições iniciais, ampliando o escopo de possíveis investigações. Por fim, o desenvolvimento de outras áreas na linha de sistemas biológicos tem mostrado que as visões tradicionais empregadas no estudo do funcionamento desses sistemas é insuficiente para explicar a complexidade dos organismos vivos e a sua maneira de realizar processamento de informação. Nesse sentido, novas visões que levem em consideração os resultados mais recentes da ciência moderna devem ser exploradas.

5.1. Considerações Finais

A seguir, são resumidas as principais contribuições deste trabalho:

- Proposta de uma metodologia para a reconstrução de redes gênicas a partir de dados de expressão. Diferente das abordagens mais empregadas, o método proposto utiliza redes bayesianas contínuas e é especialmente projetado para conjuntos de dados reduzidos e bastante ruidosos. Neste sentido, a proposta é considerada inovadora, pois os conjuntos de dados de expressão gênica são em geral muito reduzidos, em relação à complexidade da tarefa de identificação de sistemas envolvida, e as técnicas tradicionais não são adequadas para trabalhar nessas condições. Essa capacidade de lidar com recursos limitados é atingida por meio de um novo método

de estimação de densidade para domínios contínuos, que dá prioridade à generalização, ao invés de especificidade quando os dados disponíveis são limitados.

- Proposta de um modelo conexionista para redes gênicas, e uma metodologia evolutiva de síntese de redes que são capazes de resolver tarefas dinâmicas. O conjunto "modelo" mais "procedimento evolutivo" conduz às chamadas redes gênicas artificiais, e a abordagem se aproxima bastante do formalismo conexionista de redes neurais artificiais, embora possua características particulares que a diferenciam deste. A proposta de modelagem conexionista é inovadora na linha de modelagem de redes gênicas. As redes gênicas obtidas para a resolução do problema de quimiotaxia virtual foram analisadas considerando a relação entre a dinâmica e estrutura, mostrando que essas duas características devem ser consideradas em conjunto. As redes gênicas artificiais, da forma como foram propostas aqui, apresentaram um grande potencial a ser explorado tanto como ferramenta de resolução de problemas como laboratório virtual para o estudo do funcionamento das redes de organismos naturais.
- Proposta de uma nova ótica para o estudo das redes gênicas, na qual o processamento de informação celular é realizado por meio de um conjunto de osciladores acoplados em diferentes modos de sincronia. Estudos em neurociência têm mostrado que o formalismo clássico de redes neurais artificiais não é adequado para explicar como se dá o processo de coordenação no cérebro. Como as redes gênicas artificiais se assemelham bastante às redes neurais artificiais, é possível esperar que elas também sejam insuficientes para explicar efeitos coordenados na célula. Sendo assim, é necessário explorar possibilidades alternativas de investigação, e as novas evidências e teorias que têm sido empregadas no estudo do cérebro podem ser de grande ajuda nesse processo. Nesse sentido, a proposta de discussão apresentada é inovadora, se diferenciando significativamente das linhas tradicionais de estudo em redes gênicas.

5.2. Perspectivas Futuras

Todas as propostas apresentadas nesta dissertação abrem muitas possibilidades de investigação futura. Para a ferramenta de inferência de redes gênicas, a extensão mais imediata é o emprego de heurísticas de busca mais eficientes no processo de otimização da estrutura das redes bayesianas. Os resultados deixaram claro que o algoritmo *Hill climbing* não é adequado para essa tarefa, mesmo contrariando o que tem sido afirmado na literatura, e deve ser possível melhorar o desempenho da ferramenta proposta através do uso de métodos de busca capazes de evitar mínimos locais.

No caso das redes gênicas artificiais, muito ainda pode ser feito. Extensões simples, como adicionar mais reações ao modelo, podem ser realizadas sem grande dificuldade. Outra possibilidade interessante é tentar evoluir redes gênicas para problemas mais realistas. O próprio caso da quimiotaxia pode ser estudado, utilizando-se uma formulação mais real para o problema. Dessa forma, é possível comparar diversas estruturas alternativas com a estrutura de quimiotaxia conhecida em bactérias reais, e tentar extrair características essenciais do sistema. É possível também explorar a técnica como ferramenta de engenharia para a resolução de problemas. As redes gênicas artificiais podem fundar um novo campo na linha de aprendizado de máquina (ao lado das redes neurais artificiais e sistemas imunológicos artificiais, por exemplo), e serem empregadas em tarefas de controle, robótica autônoma, ou até como técnicas de mineração de dados, como em clusterização, regressão e predição de séries temporais.

Por fim, a proposta de uma rede gênica como um conjunto de osciladores coordenados em uma estrutura fractal traz uma perspectiva inteiramente nova para o estudo das redes gênicas. A partir de conhecimentos em osciladores biológicos e conceitos de neurociência e teoria de sistemas complexos já formalizados, foi possível promover uma visão bastante ampla com algumas hipóteses simples, indicando que a idéia se mostra promissora. Embora o estudo na linha de osciladores biológicos acoplados esteja em ascensão, ainda não houve um debate consistente sobre as possíveis formas de codificação da informação em uma rede gênica, assim como há na neurociência sobre a codificação da informação no cérebro, na qual o comportamento oscilatório tem papel determinante. Logo, a proposta apresentada no Capítulo 4 pode ser interpretada como uma tentativa inicial de promover esse debate.

Referências

1. **(ABBOTT & VAN VREESWIJK, 1993)** Abbott, L. F. & van Vreeswijk, C. (1993). Asynchronous states in networks of pulse-coupled oscillators, *Physical Rev. E*, 48:1483-1490.

2. **(ADRIAN, 1926)** Adrian, A.D. (1926). The impulses produced by sensory nerve endings: Part I, *J. Physiol*. (Lond.), 61:49-72.

3. **(AGUIRRE, 2004)** Aguirre, L. A. (2004). *Introdução à identificação de sistemas: técnicas lineares e não-lineares aplicadas a sistemas reais*. 2ª edição, Belo Horizonte, Editora UFMG.

4. **(AKAIKE, 1974)** Akaike, H. (1974). A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, vol. AC-19, pp.716-23, 1974.

5. **(ALBERT, *et al*., 2000)** Albert, R., Jeong, H. & Barabasi, A.-L, (2000). Error and attack tolerance in complex networks, *Nature*, 406:387-482.

6. **(ALBERTS *et al*., 1989)** Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J.D. (1989). *Molecular Biology of the Cell*. Garland, New York.

7. **(ARKIN *et al*., 1998)** Arkin, A., Ross, J., & McAdams, H.A. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage-infected *Escherichia coli* cells, *Genetics* 149:1633-1648.

8. **(ARKIN, 1998)** Arkin, R.C. (1998). *Behavior-Based Robotics*. The MIT Press, Cambridge, MA, EUA.

9. (**BALDI & BRUNAK, 2001)** Baldi, P. & Brunak, S., (2001), Bioinformatics - The Machine Learning Approach, 2nd Ed., MIT Press, Cambridge, Massachussetts.

10. **(BAK, 1997)** Bak, P. (1997). *How Nature Works*, Oxford University Press, 1997.

11. **(BARABÁSI, 2002)** Barabási, A.-L. (2002). *Linked: The New Science of Networks*. Perseus Publishing, Cambridge, 2002.

12. **(BARKAY & LEIBLER, 1997)** Barkai N. & Leibler S., (1997). Robustness in simple biochemical networks, *Nature*, 387: 913-917.

13. **(BELLMAN, 1961)** Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

14. **(BERTONE & GERSTEIN, 2001)** Bertone, P. & Gerstein, M. (2001). Integrative Data Mining: The New Direction in Bioinformatics – Machine learning for analyzing genome-wide expression profiles, *IEEE Engineering in Medicine and Biology*. vol. 20, pp. 33-40.

15. **(BEZERRA *et al*., 2005)** Bezerra, G. B., Barra, T. V., de Castro, L. N. & Von Zuben, F. J. (2005). Adaptive Radius Immune Algorithm for Data Clustering, Em C. Jacob, M.L. Pilat, Bentley, P.J. and J. Timmis (Eds.), *Artificial Immune Systems, Lecture notes in Computer Science*, Springer-Verlag, Berlin, vol. 3627, pp. 290-303, 2005.

16. **(BILMES, 1998)** Bilmes, J. (1988). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Tech. Rep. ICSI-TR-97-021, University of Berkeley, 1998.

17. **(BISHOP, 1995)** Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Oxford, U.K.: Oxford Univ. Press, 1995.

18. **(BODEN, 1998)** Boden, M.A. (1998). Autonomy and Artificiality, Em A. Clark and J. Toribio (Eds.), *Artificial Intelligence and Cognitive Science: Cognitive Architectures in Artificial Intelligence*, Garland Publishing, Inc., New York, EUA.

19. **(BONGARD, 2002)** Bongard, J. (2002). Evolving modular genetic regulatory networks, *Proceedings of the IEEE 2002 Congress on Evolutionary Computation*, IEEE Press 1872–1877.

20. **(BORISUK & TYSON, 1998)** Borisuk, M. T., & Tyson, J. J. (1998). Bifurcation analysis of a model of mitotic control in frog eggs, *J. Theor. Biol.* 195:69-85.

21. **(BRAGIN *et al.*, 1995)** Bragin, A. Jandó, G., Nádasdy, Z., Hetke, J.K., Wise, K. & Buzsáki, G. (1995). Gama (40-100hz) oscillation in the hippocampus of the behaving rat, *J. Neurosci*., 15:47-60.

22. **(BRUGGE & MERZENICH, 1973)** Brugge, J. F. & Merzenich, M. M. (1973). Responses of neurons in auditory cortex of the macaque monkey to monaural and binaural stimulations, *J. Neurophysiol*. 36:1138-1158.

23. **(BULLOCK *et al.*, 1990)** Bullock, T. H., Buzsaki, G. & McClune, M. C. (1990). Coherence of compound field potential reveals discontinuities in the ca1-subiculum of the hippocampus in freely-moving rats, *Neuroscience*, 38:609-619.

24. **(CAPRA, 1982)** Capra, F. (1982). *The Turning Point*, Simon & Schuster, New York, USA.

25. **(CARPENTER, 1996)** Carpenter, R. H. S. (1996). *Neurophysiology*, Arnolds, London.

26. **(CHERRY & ADLER, 2000)** Cherry, J.L. & Adler, F.R. (2000). How to make a biological switch, *J. Theor. Biol*. 203:117-133.

27. **(CHICKERING *et al.*, 2004)** Chickering, D.M., Heckerman, D. & Meek. C. (2004). Large Sample Learning of Bayesian Networks is NP-Hard, *The Journal of Machine Learning Research*, Vol 5, pp. 1287-1330.

28. **(CHRISTENDAT *et al.*, 2000)** Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxell, K.L., Wu, N., Mc-Intosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M. & Arrowsmith, C.H. (2000), Structural proteomics of an archaeon, *Nat. Struct*. Biol., vol. 7, pp. 903-908, 2000.

29. **(COOPER & HERSKOVITS, 1992)** Cooper, G. & Herskovits, E. (1992), A bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9:309-347, 1992.

30. **(DAVIES & MOORE, 2000)** Davies, S. & Moore, A. (2000). Mix-Nets: Factored Mixtures of Gaussians in Bayesian Networks with Mixed Continuous and Discrete Variables, *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, pp. 168-175, 2000.

31. **(DE CASTRO, 2006)** de Castro, L. N. (2006). *Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications*, Chapman & Hall/CRC, 2006.

32. **(DE CASTRO & TIMMIS, 2002)** de Castro, L. N. & Timmis, J. I. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer-Verlag, 2002.

33. **(DE CHARMS & MERZENICH, 1996)** de Charms, R. C. & Merzenich, M. M. (1996). Primary cortical representation of sounds by the coordination of action-potential timing, *Nature*, 381:610-613.

34. **(DE JONG, 2002)** de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review, *J. Comput Biol*. 9:67-103.

35. **(DECKARD & SAURO, 2004)** Deckard, A. & Sauro, H.M. (2004). Preliminary studies on the in silico evolution of biochemical networks, *Chembiochem*, 5(10):1423-3.

36. **(ENGEL *et al*., 1991)** Engel, A. K., Konig, P., Kreiter, A., & Singer, W. (1991). Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science*, 252:1177-1179.

37. **(ENGEL *et al*., 1997)** Engel, A. K., Roelfsema, P. R., Fries, P., Brecht, M. & Singer, W. (1997). Role of the temporal domains for response selection and perceptual binding, *Cerebral Cortex*, 7:571-582.

38. **(FORSTER, 2000)** Forster, M. R. (2000). Key Concepts in Model Selection: Performance and Generalizability, *Journal of Mathematical Psychology*, 44, 205-231, 2000.

39. **(FRANÇOIS & HAKIM, 2004)** François, P. & Hakim, V. (2004). Design of genetic networks with specified functions by evolution *in silico*, *Proc. Natl. Acad. Sci*., Jan 13;101(2):580-5, USA.

40. **(FRIEDMAN *et al*., 1999)** Friedman, N., Linial, M., Nachman & I., Pe'er, D. (1999). Using Bayesian Networks to Analyze Expression Data, *Proc. of the 4th annual Inter. Conference on Comp. Mol. Biology*, pp. 127–135, Tokyo, Japão 1999.

41. **(FRIEDMAN *et al*., 1998)** Friedman, N., Murphy, K. & Russel, S. (1998). Learning the structure of dynamic probabilistic networks, *Proc. of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, 1998.

42. **(FRIESEN & BLOCK, 1984)** Friesen, W. O. & Block, G. D. (1984). What is a biological oscillator?, *Am J Physiol Regul Integr Comp Physiol*, 246: R847-R853, 1984.

43. **(GARCIA-OJALVO *et al.*, 2004)** Garcia-Ojalvo, J., Elowitz, M. B. & Strogatz, S. H. (2004). Modeling a synthetic multicellular clock: Repressilators coupled by quorum sensing, *PNAS*, July 27, 2004; 101(30): 10955 - 10960.

44. **(GEARD, 2004)** Geard, N. (2004). "Modelling Gene Regulatory Networks: Systems Biology to Complex Systems", ACCS Draft Technical Report, 2004.

45. **(GEMAN *et al.*, 1992**) Geman, S., Bienenstock, E. & Doursat, R. (1992). Neural networks and the bias/variance dilemma, *Neural Computation*, vol. 4, no. 1, pp. 1-58, 1992.

46. **(GERSTNER *et al.*, 1993)** Gerstner, W., Ritz, R. & van Hemmen, J. L. (1993). A biologically motivated and analytically soluble model of collective oscillations in the cortex I. Theory of weak locking, *Biol. Cybern.*, 68:363-374.

47. **(GIBSON & MJOLSNESS, 2001)** Gibson, M.A. & Mjolsness, E. (2001). Modeling the activity of single genes. Em J.M. Bower & H. Bolouri, (eds). *Computational Modeling of Genetic and Biochemical Networks*, 1–48. MIT Press, Cambridge, MA, EUA.

48. **(GILLESPIE, 1977)** Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem*. 81(25), 2340–2361.

49. **(GIROSI *et al.*, 1995)** Girosi, F., Jones, M., Poggio, T. (1995). Regularization Theory and Neural Networks Architectures. *Neural Computation*, vol. 7, no. 2, pp. 219-269, 1995.

50. **(GOLDBETER, 1996)** Goldbeter, A. (1996). *Biochemical Oscillations and Cellular Rhythms*, Cambridge Univ. Press, Cambridge, EUA, 1996.

51. **(GOLDENFELD, 1992)** Goldenfeld, N. (1992). *Lectures on Phase Transitions and the Renormalization Group*, Perseus Publishing (1992).

52. **(GOLOMB *et al.*, 1992)** Golomb, D., Hansel, D., Shraiman, S. & Sompolinsky, H. (1992). Clustering in globally coupled phase oscillators, *Physical Rev. A.*, 45:3516-3530.

53. (**GOLUB *et al.*, 1999**) Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Merisov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S. (1999). Molecular Classification of Cancer: class discover and class prediction by gene expression monitoring, *Science*, Vol. 286. no. 5439, pp. 531 – 537.

54. **(GONZE *et al.*, 2005)** Gonze, D., Bernard, S., Waltermann, C., Kramer, A. & Herzel, H. (2005). Spontaneous Synchronization of Coupled Circadian Oscillators, *Biophys. J.*, July 1, 2005; 89(1):120-129.

55. **(GOODWIN, 1963)** Goodwin, B.C. (1963). *Temporal Organization in Cells*, Academic Press, New York, EUA.

56. (**GOODWIN, 1965**) Goodwin, B.C. (1965). Oscillatory behavior in enzymatic control processes. Em G. Weber, ed. *Advances in Enzyme Regulation*, 425–438. Pergamon Press, Oxford.

57. **(GRANNAN *et al.*, 1993)** Grannan, E. R., Kleinfeld, D. & Sompolinsky, H. (1993). Stimulus-dependent synchronization of neuronal assemblies, *Neural Computation*, 5:550-569.

58. **(GRAY AND SINGER, 1989)** Gray, C. M. & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex, *Proc. Nat. Acad. Sci*., 86:1698-1702.

59. (**HAKEN, 1983)** Haken, H. (1983). *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, 3rd rev. enl. ed. New York: Springer-Verlag, 1983.

60. **(HALLINAN & WILES, 2004a)** Hallinan, J. & Wiles, J. (2004). Evolving genetic regulatory networks using an artificial genome, In Chen, Y.P.P., eds., *Second Asia-Pacific Bioinformatics Conference (APBC2004)*, Volume 29 of CRPIT., Dunedin, New Zealand, ACS 291–296.

61. (**HALLINAN & WILES, 2004b)** Hallinan, J. & Wiles, J. (2004). Asynchronous dynamics of an artificial genetic regulatory network, *Ninth International Conference on the Simulation and Synthesis of Living Systems (ALife9)* Boston, September 12 - 15.

62. (**HALLINAN, 2004)** Hallinan, J. (2004). Gene duplication and hierarchical modularity in intracellular interaction networks, *BioSystems* 74(1-3):51- 62.

63. **(HAMMOND, 1993)** Hammond, B. J. (1993). Quantitative study of the control of HIV-1 gene expression, *J. Theor. Biol*., 163:199-221.

64. **(HAYKIN, 1994)** Haykin, S. (1994*). Neural Networks: A Comprehensive Foundation*, IEEE Press/Macmillan College Publishing Company, New York, USA.

65. **(HECKERMAN, 1997)** Heckerman, D. (1997). A Bayesian Approach to Causal Discovery, Technical Report MSR-TR-97-05, 1997.

66. **(HESS, 1979)** Hess, B. (1979). The glycolytic oscillator. *J. Exp. Biol.*, 81:7-14, 1979.

67. **(HOFMANN & TRESP, 1996)** Hofmann, R. & Tresp, V. (1996). Discovering Structure in Continuous Variables Using Bayesian Networks, *Advances in Neural Information Processing Systems*, MIT Press, 1996

68. **(HOLLAND, 1998)** Holland, J. H. (1998). *Emergence: From chaos to order*, Helix Books: Reading, MA, EUA.

69. **(HOPFIELD & HERTZ, 1995)** Hopfield, J. J. & Herz, A. V. M. (1995). Rapid local synchronization of action potentials: toward computation with coupled integrate-and-fire neurons, *Proc. Natl. Acad. Sci*., USA, 92:6655-6662.

70. **(HUBEL & WIESEL, 1962)** Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields of single neurons in the cat's striate cortex, *J. Physiol*., 148:574-591.

71. **(JACOB, 1998)** Jacob, F. (1998). *Of Flies, Mice, and Men*, Harvard University Press, Cambridge MA, USA.

72. **(JACQUETTE, 1994)** Jacquette, D. (1944). *Ockham's Razor. Philosophy of Mind*, Engleswoods Cliffs, N.J., Prentice Hall, pp. 34-36, 1994.

73. **(JEONG *et al*., 2000)** Jeong, H., Tombor, B., Albert, A., Oltvai, Z.N. & Barabási. A.-L., (2000). The large-scale organization of metabolic networks, *Nature*, (407):651-654.

74. **(JEONG *et al*., 2001)** Jeong, H., Mason, S., Barabási, A. -L. & Oltvai, Z. N. (2001). Centrality and lethality of protein networks, *Nature*, vol. 411, pp. 41-2, 2001.

75. **(KAUFFMAN, 1993)** Kauffman, S. (1993). *The Origins of Order*, Oxford University Press.

76. **(KELLER, 1994)** Keller, A.D. (1994). Specifying epigenetic states with autoregulatory transcription factors, *J. Theor. Biol.* 170, 175–181.

77. **(KELSO, 1995)** Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-organization of Brain and Behavior*, Cambridge, MA: The MIT Press, 1995.

78. **(KHAN *et al*., 2002)** Khan, R., Zeng, Y., Garcia-Frias, J. & Gao, G. (2002). A Bayesian Modeling Framework for Genetic Regulation, CSB, pp.330-332, 2002.

79. **(KUO *et al*., 2004)** Kuo, P.D., Lieier, A. & Banzhaf, W. (2004). Evolving Dynamics in an Artificial Regulatory Network Model, Proc. of the Parallel Problem Solving, Em *Nature Conference (PPSN-04)*, Birmingham, UK, September 2004, Yao X., Burke E., Lozano J.A., Smith J., Merelo-Guervós J.J., Bullinaria J.A., Rowe J., Tino P., Kabán A., Schwefel H.-P. (Eds.), Springer, LNCS 3242, Berlin, pp. 571 – 580.

80. **(KURAMOTO, 1990)** Kuramoto, Y. (1990). Collective synchronization of pulse-coupled oscillators and excitable units, *Physica D*, 50:15-30.

81. **(LI *et al*., 2006)** Li, C., Chen, L. & Aihara, K. (2006). Synchronization of coupled nonidentical genetic oscillators, *Phys. Biol*. 3:37-44.

82. **(MAASS & BISHOP, 1999)** Maass, W. & Bishop, C. M., Eds. (1999). *Pulsed Neural Networks*, MIT Press, Cambridge, Mass.

83. **(MCADAMS & ARKIN, 1997)** McAdams H. H. & Arkin, A. (1997). Stochastic mechanisms in gene expression, Em *Proceedings of the National Academy of Sciences of the USA*, vol. 94, pp 814-819. National Academy of Sciences, 1997.

84. **(MACADAMS & SHAPIRO, 1995)** McAdams, H. H. & Shapiro, L. (1995). Circuit simulation of genetic networks, *Science*, 269, 650–656.

85. **(MACKAY & MCCULLOCH, 1952)** MacKay, D. & McCulloch, W. S. (1952). The limiting information capacity of a neuronal link, Bull. Math. Biophys. 14:127-135.

86. (**MAHAFFY, 1984)** Mahaffy, J. M. (1984). Cellular control models with linked positive and negative feedback and delays: I. *The models. J. Math. Biol.* 106, 89–102.

87. **(MEECH, 1979)** Meech, R. W. (1979). Membrane potential oscillations in molluscan "burster" neurons, *Exp. Biol.* 81:93-112, 1979.

88. **(MIROLLO & STROGATZ, 1990)** Mirollo, R. E. & Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM. J. Appl. Math.*, 50:1645-1662.

89. **(NEUENSCHWANDER *et al.*, 1996)** Neuenschwander, S., Engel, A.K., Konig, P., Singer, W. & Varela, F. J. (1996). Synchronization of neuronal responses in the optic tectum of awake pigeons. Vis. Neurosci., 13:575-584.

90. **(NOLFI & FLOREANO, 2002)** Nolfi, S. & Floreano, D., (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*, The MIT Press, Cambridge, MA, USA.

91. **(O'KEEFE & BURGESS, 1996)** O'Keefe, J. & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons, *Nature*, 381:425-428.

92. **(ONG *et al.*, 2002)** Ong, I. M., Glasner, J.D. & Page, D. (2002). Modeling regulatory pathways in E. coli from time series expression profiles, *Bioinformatics*, vol. 18, pp. 241-8, 2002.

93. **(PAN *et al.*, 2003)** Pan, W., Len, J. & Le, C. T. (2003). A mixture model approach to detecting differentially expressed genes with microarray data, *Funct. Integr. Genomics*, vol. 3, pp.117-124, 2003.

94. **(PE'ER *et al.*, 2001)** Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles, *Bioinformatics*, vol. 17, pp. 215-224, 2001.

95. **(PEARL, 1988)** Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, Calif, 1988.

96. **(PEÑA, 2004)** Peña, J. M. (2004). Learning and Validating Bayesian Network Models of Genetic Regulatory Networks, Em *Proceedings of the Second European Workshop on Probabilistic Graphical Models*, 161-168, 2004.

97. **(PERRIN *et al.*, 2003)** Perrin, B.E. *et al*. (2003). Gene Networks Inference Using Dynamic Bayesian Networks, *Bioinformatics*, vol. 19, pp. 138-148, 2003.

98. **(PRECHTL *et al.*, 1997)** Prechtl, J., Cohen, L. B., Pesaran, B., Mitra, P. P. & Kleinfield, D. (1997). Visual stimuli induce waves of electrical activity in turtle cortex, *Proc. Nat. Acad. Sci.*, 94:7621-7626.

99. **(PRIGOGINE & STENGERS, 1984)** Prigogine, I. & Stengers, I. (1984). *Order out of chaos*, Bantam Books, New York, USA.

100. **(RAVASZ *et al.*, 2002)** Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., & Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks, *Science*, Vol. 297, pp. 1551-1555.

101. **(REIL, 1999)** Reil, T. (1999). Dynamics of gene expression in an artificial genome: Implications for biological and artificial ontogeny, Em Floreano, D., Nicoud, J.D., Mondada, F., eds., *Advances in Artificial Life – Proceedings of the 5th European Conference on Artificial Life (ECAL)*. Volume 1674 of Lecture Notes in Computer Science., Springer–Verlag 457–466.

102. **(REINITZ & SHARP, 1995)** Reinitz, J. & Sharp, D. H. (1995). Mechanism of eve stripe formation, *Mech. Dev*., 49:133-158.

103. **(REINITZ & VAISNYS, 1990)** Reinitz, J. & Vaisnys, J. R. (1990). Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of co-operativity, *J. Theor. Biol.* 145, 295–318.

104. (**RENSBERGER, 1996)** Rensberger, B. (1996). *Life Itself: Exploring the realm of the living cell*, Oxford University Press, Oxford, USA.

105. **(RICE *et al*., 2004)** Rice, J. J., Tu, Y. & Stolovitzky, G. (2004). Reconstructing biological networks using conditional correlation analysis, *Bioinformatics*, vol. 21, n. 6, pp. 765-773, 2004.

106. **(RIEHLE *et al*., 1997)** Riehle, A. Grun, S., Diesmann, M. & Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function, *Science*, 278:1950-1953.

107. **(ROSS-MACDONALD *et al*., 1999)** Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F.K., Iwasaki, H., Hager, K., Gerstein M., Miller, P., Roeder, G.S. & Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption, *Nature*, vol. 402, pp.413-418.

108. **(SCHNEIDER & KAY, 1994)** Schneider, E. D., Kay, K. J. (1994), Life as a Manifestation of the Second Law of Thermodynamics, *Mathematical and Computer Modeling*, Vol. 19, No. 6-8, pp. 25-48.

109. **(SCHWARZ, 1978)** Schwarz, G. (1978). Estimating the Dimension of a Model, *Annals Statistics*, vol. 6, pp.461-5, 1978.

110. **(SCOTT, 1992)** Scott, D.W. (1992). *Multivariate Density Estimation*, NY: Wiley, 1992.

111. **(SINGER & GRAY, 1995)** Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis, *Ann. Rev. Neurosci*., 18:555-586.

112. **(SKAGGS *et al*., 1996)** Skaggs, W. E., McNaughton, B.L., Wilson, M.S. & Barnes, C. A. (1996). Theta-phase precession in hippocampal neuronal populations and the compression of temporal sequences, *Hippocampus*, 6:149-172.

113. **(SMITH *et al*., 2003)** Smith, V. A., Jarvis, E. D. & Hartemink, A. J. (2003). Influence of network topology and data collection on network inference, *Pac. Symp. Biocomput*., 164–175, 2003.

114. **(SMOLEN *et al*., 2000)** Smolen, P., Baxter, D. A., & Byrne, J. H. (2000). Modeling transcriptional control in gene networks: Methods, recent results, and future directions. *Bull. Math. Biol*. 62, 247–292.

115. **(SOFTKY & KOCH, 1993)** Softky, W. R. & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neuroscience*., 13:334-350.

116. **(SOMOGYI & SNIEGOSKY, 1996)** Somogyi, R. & Sniegoski, C. A. (1996). Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation. *Complexity* 1(6), 45–63.

117. **(SONG *et al*., 2005)** Song, C., Havlin, S. & Makse, H. A. (2005), Self-similarity of complex networks, *Nature*, 433, 392-395.

118. **(SPIRTES *et al*., 2000)** Spirtes, P., Glymour, C. & Scheines, R. (2000). Constructing Bayesian network models of gene expression networks from microarray data, Em *Proc. of the Atlantic Symp. on Comp. Biol., Genome Inf. Syst. and Tech*., 2000.

119. **(STROGATZ & STEWART, 1993)** Strogatz, S. H. & Stewart, I. (1993). Coupled oscillators and biological synchronization, *Scientific American*, Dec. 93, 68-75.

120. **(STROGATZ, 2003)** Strogatz, S. (2003). *Sync: The emerging science of spontaneous order*, Hyperion Books, New York, USA.

121. **(SZALLASI & LIANG, 1998)** Szallasi, Z., & Liang, S. (1998). Modeling the normal and neoplastic cell cycle with 'realistic Boolean genetic networks': Their application for understanding carcinogenesis and assessing therapeutic strategies. Em R.B. Altman, A.K. Dunker, L. Hunter, & T.E. Klein, eds. *Proc. Pac. Symp. Biocomput. (PSB'98)*, vol. 3, 66–76, Singapore, World Scientific Publishing.

122. **(TAKAHASHI & ZATZ, 1982)** Takahashi, J. S. & Zatz, M. (1982). Regulation of circadian rhythmicity, *Science*, 217:1104–1111.

123. **(TERMAN & WANG, 1995)** Terman, D. & Wang, D. L. (1995). Global competition and local cooperation in a network of neural oscillators, *Physica D*, 81:148-176.

124. **(THOMAS, 1998)** Thomas, R. (1998). Laws for the dynamics of regulatory networks. *Int. J. Dev. Biol*. 42, 479–485.

125. **(UERTZ *et al*., 2000)** Uertz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadomodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J.M. (2000), A comprehensive analysis of protein-protein interactions in *Saccaromyces cerevisiae*, *Nature*, vol. 402, pp.413-418.

126. **(USHER *et al.*, 1993)** Usher, M., Schuster, H. S. & Niebur, E. (1993). Dynamics of populations of integrate-and-fire neurons, partial synchronization and Memory, *Neural Computation*, 5:570-586.

127. **(VAADIA *et al.*, 1995)** Vaadia, E., Haalman, I., Abeles, M., Bergman, H., Prut, Y., Slovin, H., & Aertsen, A. (1995). Dynamics of neuronal interactions in monkey cortex in relation to behavioural events, *Nature*, 373:515-518.

128. **(VAN BERLO *et al.*, 2003)** Van Berlo, R. J. P., van Someren, E. P. & Reinders, M. J. T. (2003). Studying the Conditions for Learning Dynamic Bayesian Networks to Discover Genetic Regulatory Networks, *Simulation*, vol. 79, Issue 12, pp. 689-702, 2003.

129. **(VELCULESCU *et al.*, 1997)** Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E., Heiter, P., Volgelstein, B. & Kinzler, K.W. (1997). Characterization of the yeast transcriptome, *Cell*, vol. 88, pp. 243-251.

130. **(VILAR *et al.*, 2002)** Vilar, J. M., Kueh, H. Y., Barkai, N. & Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators, *Proc Natl Acad Sci*, 99:5988-5992.

131. **(VINGRON & HOJEISEL, 1999)** Vingron M. & Hojeisel, J. (1999). Computational aspects of expression data, *J. Mol. Med*., vol. 77, pp. 3-7, 1999.

132. **(WATTS, 1999)** Watts, D. J. (1999). *Small Worlds: The dynamic of networks between order and chaos*, Princeton University Press, Princeton, New Jersey, EUA.

133. **(WEAVER *et al.*, 1999)** Weaver, D. C., Workman, C. T. & Stormo, G. D. (1999). Modeling regulatory networks with weight matrices, Em *Pacific Symposium on Biocomputing*, vol. 4, pp. 112-23, 1999.

134. **(WEISBUCH, 1986)** Weisbuch, G. (1986). Networks of automata and biological organization, *J. Theor. Biol.* 121, 255–267.

135. **(YU *et al.*, 2004)** J. Yu, *et al*. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data, *Bionformatics*, vol. 20, no. 18, pp 3594-3603, 2004.

136. **(ZHU *et al.*, 2000)** Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., e Snyder, M. (2000), Analysis of yeast protein kinases using protein chips, *Nat. Genet.*, vol. 26, pp. 283-289.

137. **(ZOU & CONZEN, 2005)** Zou, M. & Conzen, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data, *Bionformatics*, vol. 21, no. 1, pp.71-79, 2005.

Apêndice

Análise Experimental das Redes Bayesianas

Este apêndice apresenta os resultados de testes experimentais realizados com redes bayesianas na tarefa de aprendizado de estrutura utilizando uma variação do algoritmo K2 (COOPER & HERSKOVITS, 1992) como heurística de busca e a máxima verossimilhança como critério de avaliação. Todos os testes se basearam na capacidade do sistema em reproduzir uma rede bayesiana pré-definida através de amostragens produzidas por esta rede. Em outras palavras, uma rede bayesiana é utilizada como modelo para produzir uma determinada amostragem e, através desta amostragem, o sistema de inferência deve ser capaz de reproduzir a mesma rede. A análise realizada aqui tem múltiplos propósitos:

1) Explorar o potencial das redes bayesianas como ferramentas de identificação de sistemas;
2) Avaliar o papel do fator representatividade/quantidade de dados no processo de inferência;
3) Analisar o impacto da complexidade dos modelos a serem inferidos na tarefa de recuperação das redes originais;
4) Investigar a influência das limitações da heurística de busca e do critério de avaliação escolhidos na qualidade da inferência.

Como resultado dessas análises, espera-se adquirir uma noção intuitiva e prática do potencial das redes bayesianas na tarefa de inferência de estrutura de redes e das dificuldades que podem ser encontradas ao longo desse processo. Para isso, a configuração do algoritmo para os experimentos realizados será propositadamente padrão, isto é, não serão consideradas maiores sofisticações relativas à heurística de busca e ao critério de avaliação empregados.

As Seções A.2.1 e A.2.2 deste apêndice trazem introduções sobre o algoritmo K2 e o critério de máxima verossimilhança, respectivamente. A Seção A.2.3 avalia a capacidade do sistema em reproduzir a rede original em função da quantidade de amostras disponíveis e a Seção A.2.4 investiga o potencial do algoritmo de busca para encontrar a distribuição

observada nas amostras, isto é, a sua capacidade de maximizar a verossimilhança. Por fim, a Seção A.2.5 apresenta um balanço das conclusões obtidas ao longo dos experimentos e discorre sobre a utilidade prática das redes bayesianas como ferramentas de inferência de relações causais entre variáveis.

A. 2.1 Heurística de Busca

O algoritmo K2 (COOPER & HERSKOVITS, 1992) de inferência de redes bayesianas funciona de forma bastante simples. Ele faz uma busca "gulosa" no espaço de possíveis estruturas de rede, à procura daquela que maximiza um determinado critério de qualidade (no caso, a verossimilhança, que será discutida na próxima seção).

Na variação do algoritmo considerada aqui (e também empregada em FRIEDMAN *et al*. (1999)), inicia-se com uma rede sem conexões, isto é, consideram-se as variáveis totalmente independentes umas das outras, e avalia-se a qualidade da rede em relação a uma dada amostragem. O próximo passo consiste em adicionar um arco à estrutura. Testam-se todas as possíveis estruturas que contêm apenas um arco, avaliando cada uma, e armazenando aquela que maximiza o critério de qualidade. Se a rede com uma conexão apresentar maior qualidade que a rede sem conexões, a nova rede substitui a anterior. A partir daí, o processo se repete considerando agora redes com duas conexões. Se a rede com duas conexões for melhor que a rede com uma conexão, aquela substitui esta. E assim sucessivamente, até que uma rede com uma conexão a mais não seja capaz de aumentar o valor do critério de qualidade. Fica-se com a rede anterior, de maior qualidade, e a busca é finalizada.

A.2.2 Verossimilhança como Critério de Qualidade

A verossimilhança é uma medida estatística que estima a probabilidade de um determinado modelo reproduzir um conjunto de amostras observado. Ou seja, ela mede o quanto a densidade de probabilidade representada pelo modelo se aproxima da distribuição apresentada nos dados. Esta é uma medida bastante utilizada como critério de seleção de modelos (uma rede bayesiana é um modelo) quando não se possui nenhum conhecimento a priori, isto é, a única informação disponível a respeito do problema são as amostras.

Porém, a verossimilhança possui algumas desvantagens. Primeiramente, não há compromisso com a manutenção de simplicidade; muito pelo contrário, ela vai exatamente contra o princípio da "navalha de Occam" (JACQUETT, 1994). Entre dois modelos que expliquem os dados de maneira semelhante (isto é, com verossimilhanças aproximadamente iguais), o critério de máxima verossimilhança tenderá a escolher sempre aquele modelo de maior complexidade.

Como conseqüência disso, vem o segundo problema: o modelo se torna excessivamente susceptível à qualidade do conjunto amostral. Segundo a máxima verossimilhança, o modelo deve possuir quantas variáveis forem necessárias para melhor se adequar aos dados observados. Isto, porém, o torna muito específico para aqueles dados. Caso as amostras se distanciem ligeiramente da distribuição verdadeira – e isto geralmente vai ocorrer – a capacidade de previsão do modelo se torna bastante comprometida. O modelo se torna pouco tolerante ao ruído inerente à característica probabilística da amostragem. Se ganha em especificidade, mas perde-se muito em generalidade.

Em outras palavras, a máxima verossimilhança evidencia um dilema ingrato envolvendo seleção de modelos: ao aumentar a especificidade do modelo, reduzindo assim o *bias*, o critério termina por aumentar, como conseqüência inevitável, a sua susceptibilidade ao ruído (variância). (Veja *bias* × *variance dilemma* em FORSTER (2000) e GEMAN *et al*. (1992).)

Uma solução mais adequada seria escolher um modelo cuja complexidade representa o ponto ótimo entre *bias* e variância, isto é, um ponto onde não é possível reduzir um sem aumentar o outro. Esta discussão, no entanto, não é o foco principal deste apêndice. O leitor interessado deve se referir à literatura sobre seleção de modelos, onde esta questão é bastante debatida (FORSTER, 2000).

É possível encontrar na literatura critérios que procuram amenizar o problema da máxima verossimilhança. Os critérios BIC (*Bayesian Information Criterion*) (SCHWARTZ, 1978) e AIC (*Akaike Information Criterion*) (AKAIKE, 1974), por exemplo, são medidas de qualidade bastante adotadas que introduzem um coeficiente de penalização da complexidade em conjunto com a verossimilhança no cálculo da qualidade do modelo. O resultado geralmente é mais interessante na prática do que o obtido com a máxima verossimilhança.

A.2.3 Descobrindo a Estrutura da Rede Original

Este experimento consiste em avaliar a capacidade do sistema em descobrir a estrutura original de uma rede bayesiana em função do tamanho da amostragem. É importante observar que o tamanho do conjunto amostral em si não é a variável mais relevante aqui. O objetivo principal é avaliar o potencial da metodologia empregada em função do nível de representatividade dos dados. Entretanto, dada a característica probabilística das amostras, a maneira mais direta de se obter amostras de maior representatividade é, logicamente, aumentando o número de amostras. Quanto maior o tamanho da amostragem, maior tende a ser a sua representatividade, de maneira assintótica. Assim, tendo infinitas amostras, a densidade de probabilidade dos dados é exatamente a densidade de probabilidade do modelo original.

Experimento 1: Heckerman *et al*. (1997)

Este experimento é reproduzido de (HECKERMAN, 1997). Ele evidencia de forma bastante ilustrativa a dependência da rede obtida em relação ao tamanho do conjunto de dados. Partindo-se da rede da Figura 6.1, onde são mostradas também as tabelas de probabilidade condicional de cada variável, foram geradas amostras a serem apresentadas ao algoritmo K2. A figura ilustra as variáveis v_1 e v_2 como binárias e independentes, e seus valores são determinados de acordo com suas respectivas tabelas de probabilidade condicional. Já a variável v_3 é uma variável dependente. Seu valor é determinado após v_1 e v_2 serem dados, e de acordo com a sua tabela de probabilidade condicional. Por exemplo, se v_1 for 1 e v_2 também for 1, então v_3 terá probabilidade 0,190 de assumir 1 e 0,810 de assumir 2. Sendo assim, a probabilidade de se obter, por exemplo, $v_1 = 1$, $v_2 = 1$ e $v_3 = 2$ é $p(v_1 = 1) \times p(v_2 = 1) \times p(v_3 = 2) = 0{,}660 \times 0{,}430 \times 0{,}810 = 0{,}230$.

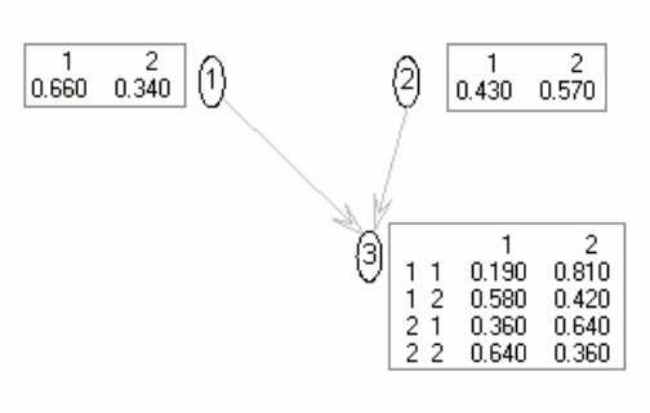


Figura 6.1 Rede bayesiana utilizada como modelo no experimento 1. Exemplo reproduzido de HECKERMAN (1997).

Como salientado anteriormente, o objetivo é observar se o sistema consegue convergir para a rede original partindo apenas dos dados. Cinco casos foram testados: 150, 250, 500, 1000 e 2000 amostras. Para cada situação, 20 conjuntos diferentes com o mesmo número de amostras foram gerados. Os resultados são mostrados na Tabela 6.1. Apenas a relação entre as variáveis 1 e 3 foi avaliada no experimento, pois a relação entre as variáveis 2 e 3 é identificada corretamente pelo sistema com facilidade.

Tabela 6.1 Resultados do experimento 1. A tabela mostra as probabilidades de a variável v_1 causar v_3 e de v_1 e v_3 estarem relacionadas após 20 execuções do algoritmo K2 para cada situação.

Nº de amostras	**p(v_1 causa v_3)**	**p(v_1 causa v_3 ou v_3 causa v_1)**
150	0,05	0,2
250	0,15	0,4
500	0,45	0,85
1.000	0,85	1
2.000	0,85	1

Note na tabela que o desempenho da inferência varia com o número de amostras. Como discutido na Seção A.2.2, esta relação já era esperada, pois à medida que o número de amostras aumenta, mais próxima a verossimilhança se torna da verdade. Quando 500 amostras são utilizadas, é possível perceber que a relação de dependência entre as variáveis já se torna bem evidente, com 85% de probabilidade, porém não há distinção clara de que v_1 causa v_3. Apenas a partir de 1.000 amostras o algoritmo é capaz de identificar corretamente a relação de causalidade.

Este exemplo traz à tona uma questão importante. Para um problema tão simples como este, são necessárias pelo menos 1.000 amostras para descobrir a estrutura original da rede. Isso é inaceitável sob praticamente quaisquer circunstâncias em problemas reais. Quase sempre um número tão elevado de amostras em relação ao de variáveis não está disponível. O problema tende a se tornar ainda mais crítico quando o número de variáveis é aumentado. Segundo o princípio da "maldição da dimensionalidade" (BELLMAN, 1961), o número de amostras necessárias para resolver um problema deste tipo aumenta exponencialmente com o número de variáveis. Ora, esta conclusão parece simplesmente eliminar qualquer esperança de recuperar a estrutura verdadeira das relações causais em problemas complexos de mundo real, a exemplo da recuperação de redes gênicas (GEARD, 2004), onde o número de variáveis tende a ser grande e a quantidade de amostras é limitada.

No entanto, em situações em que nenhum conhecimento a priori está disponível, qualquer informação, mesmo que imprecisa, é considerada de grande relevância. Veja que com 500 amostras é possível descobrir que existe uma forte relação de causalidade entre as variáveis, mesmo que o sentido da relação não esteja definido. Infelizmente, essa condição não ajuda muito. 500 amostras é ainda um número muito alto, visto que está se considerando aqui um número reduzido de variáveis. Passa-se de uma situação "extremamente difícil" para uma "muito difícil", o que não é de grande valia.

A despeito da aparente dramaticidade da questão exposta acima, cabe lembrar que a relação entre as variáveis v_2 e v_3 é facilmente percebível pelo algoritmo, como descrito anteriormente. Mais uma vez, quando nenhum conhecimento a priori é sabido, ter certeza da relação de causalidade entre um subconjunto de variáveis pode ser considerado de extrema importância, o que faz da técnica uma ferramenta útil.

Experimento 2: Exemplo clássico da chuva

Este é um exemplo clássico da literatura. A rede bayesiana consiste de 4 variáveis binárias, onde 1 significa *não* e 2 significa *sim*. A estrutura da rede e o significado lingüístico das variáveis são mostrados na Figura 6.2. Veja que todas as variáveis são binárias e que a variável v_1 (nublado) é a única variável independente. As variáveis v_2 (regador) e v_3 (chuva) dependem apenas de v_1, e a variável v_4 (grama molhada) depende

simultaneamente de v_2 e v_3, e, como conseqüência, indiretamente de v_1 também. Sendo assim, para determinar o valor da variável v_4 é preciso saber antes todas as outras variáveis. Por exemplo, se $v_1 = 1$, v_2 tem igual probabilidade (0,5) de assumir 1 ou 2. A variável v_3, por sua vez, tem 0,8 de probabilidade de assumir 1 e 0,2 de assumir 2. Uma vez determinados v_2 e v_3, podemos determinar agora v_4. Digamos que $v_2 = 1$ e $v_3 = 2$, logo v_4 terá 0,1 de chance de assumir 1 e 0,9 de assumir 2. Em termos de significado lingüístico, se o céu está nublado ($v_1 = 2$), se eu não usei o regador ($v_2 = 1$) e se choveu ($v_3 = 2$) então a probabilidade de que a grama esteja molhada ($v_4 = 2$) é 0,9.

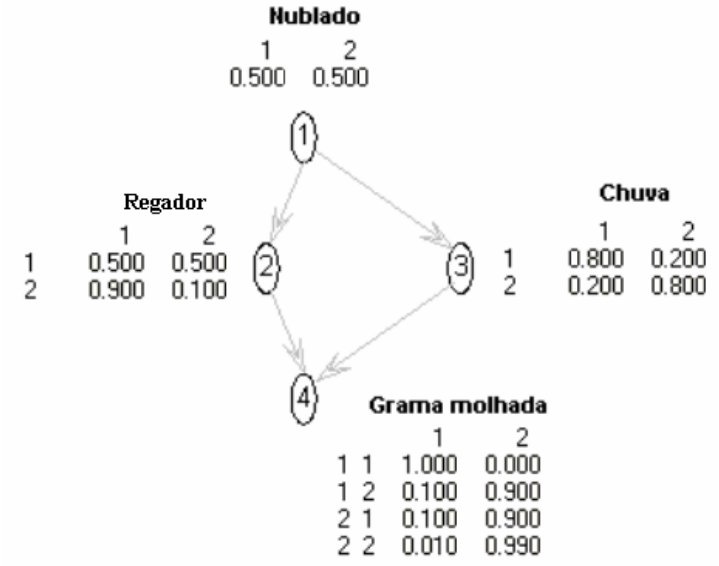
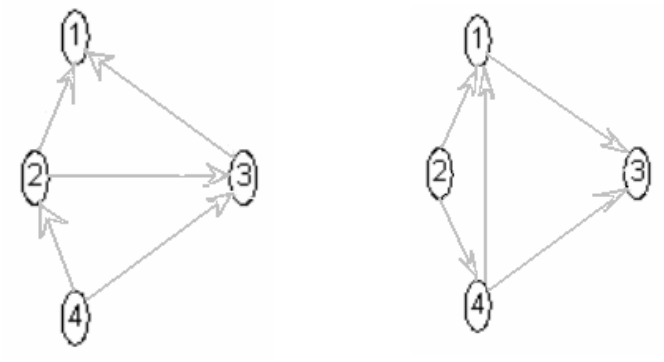


Figura 6.2 Exemplo clássico da chuva com 4 variáveis binárias. 1 significa *não* e 2 significa *sim*.

O algoritmo K2 foi utilizado para resolver o problema para 200, 1.000, 2.000, 10.000 e 50.000 instâncias. Para as 4 primeiras situações, o algoritmo oscilou entre duas estruturas, nenhuma delas exatamente a original, mostradas na Figura 6.3(a) e (b). Para 50.000 variáveis, o algoritmo encontrou apenas a estrutura mostrada na Figura 6.3(b).

O algoritmo K2 se mostrou incapaz de recuperar a estrutura original do problema, muito embora tenha sido capaz de relacionar as variáveis com certa eficiência. Veja na Figura 6.3(a) que, mesmo que o sentido das setas não esteja de acordo com o modelo original, a direção do relacionamento causal está correta, embora uma conexão adicional relacionando 2 e 3 tenha sido inserida. O mesmo acontece com a estrutura da Figura 6.3(b), sendo que a conexão adicional relaciona 1 com 4.



(a) **(b)**

Figura 6.3 (a) Estrutura encontrada para 200, 1.000, 2.000 e 10.000 amostras. (b) Estrutura encontrada em todas as situações, inclusive a com 50.000 amostras.

Para analisar os resultados obtidos, vamos considerar que a amostragem com 50.000 amostras é suficientemente grande para representar a distribuição verdadeira adequadamente, isto é, vamos considerar que, mesmo com infinitas amostras, o resultado seria o mesmo da Figura 6.3(b). Sendo assim, duas questões merecem observação especial (por conveniência, essas questões serão forçosamente tratadas separadamente aqui):

Por que o algoritmo introduziu uma conexão a mais na rede, sendo que as variáveis em questão não estão diretamente relacionadas?

Por que não foi possível determinar com exatidão o sentido das relações causais, dado que a representatividade da amostragem é elevada?

Analisaremos agora a primeira questão. A segunda será discutida nas análises do experimento 3 desta seção e na Seção A.2.4.

Uma possível explicação para o resultado destacado na questão 1 é a seguinte. Um modelo com mais variáveis pode explicar com igual ou maior precisão um fenômeno qualquer do que um modelo semelhante, mas com uma variável a menos. Se o modelo com menos variáveis explica perfeitamente o fenômeno, então o modelo com mais variáveis pode explicar perfeitamente também, basta considerar o valor da variável adicional como nulo. Diz-se que esses modelos são "modelos aninhados" (*nested models*), segundo a teoria de seleção de modelos.

Seguindo este raciocínio, agora no contexto das redes bayesianas, se uma rede com 4 arcos explica bem um conjunto de dados, uma rede com 1 ou mais arcos além desses 4

pode explicar os mesmos dados de forma igual ou melhor. Ou seja, estas redes são modelos aninhados. Como o critério de máxima verossimilhança não penaliza a complexidade, o modelo mais complexo tenderá a ser o escolhido (essa particularidade foi descrita na Seção A.2.2), sendo, portanto, esta a razão para as redes encontradas possuírem uma conexão extra.

Não se pode desconsiderar também que o algoritmo K2 pode estar realizando uma busca ineficiente, isto é, talvez a rede original, ou uma outra rede qualquer, possua uma verossimilhança maior que a da rede encontrada. Dessa forma, a explicação dada acima não se aplica necessariamente.

Experimento 3: Exemplo da gravidez

Esta rede bayesiana representa uma relação causal que determina a probabilidade de uma mulher estar grávida ou não, dado o estado de uma série de variáveis. Estes dados foram encontrados em http://www.cs.huji.ac.il/labs/compbio/Repository/. A rede possui 6 variáveis, sendo a primeira com 7 valores discretos e as outras binárias. A Figura 6.4 mostra a rede juntamente com as tabelas de probabilidade de cada variável.

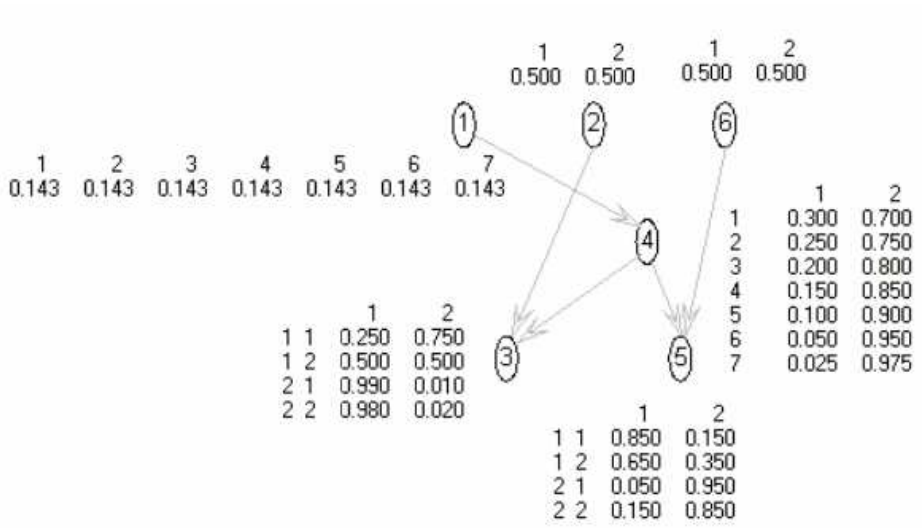
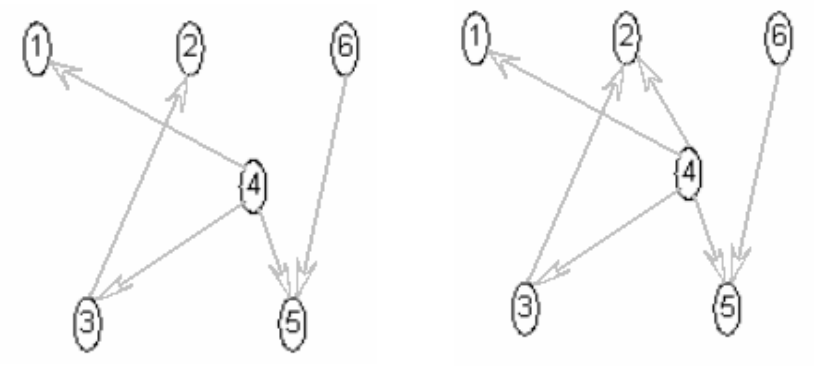


Figura 6.4 Exemplo da gravidez. Rede bayesiana com 6 variáveis, sendo a primeira com 7 valores discretos e as outras binárias.

Para amostragens com 1000 e 2000 dados, o algoritmo oscilou entre dois tipos de estruturas, mostradas nas Figura 6.5(a) e (b). A rede da Figura 6.5(a) corresponde exatamente à mesma estrutura relacional do exemplo original, sendo que o sentido de dois

arcos é diferente. Já a Figura 6.5(b) mostra uma rede igual à da Figura 6.5(a), porém com um arco a mais, correspondendo assim a um modelo aninhado. Para amostragens com 4000 e 8000 dados, apenas a estrutura da Figura 6.5(b) foi encontrada, quando não uma estrutura ainda mais complexa.



(a) **(b)**

Figura 6.5 (a) Rede encontrada com direções das relações causais semelhantes ao modelo original. (b) Rede encontrada com uma conexão adicional.

Para este problema, o algoritmo parece ter obtido um desempenho relativamente bom. Ele foi capaz de encontrar a estrutura da rede original em termos de relacionamento de variáveis, mesmo sendo esta rede mais complexa que as anteriores. Entretanto, parece que o problema da complexidade adicional provocada pela medida de qualidade da rede persiste. Este resultado reforça o fato de que uma medida que também penalize a complexidade de um modelo pode ser mais adequada que considerar simplesmente a máxima verossimilhança.

Vale ressaltar também que, assim como no experimento 2, o sentido das relações causais não pôde ser recuperado adequadamente (este resultado está relacionado à questão 2, levantada no experimento 2 da Seção A.2.3), embora as redes encontradas possuam arcos exatamente entre as mesmas variáveis. Redes deste tipo, com conexões entre as mesmas variáveis não importando o sentido, são ditas equivalentes de Markov (HECKERMAN, 1997). Embora não seja uma relação universal, redes equivalentes de Markov muitas vezes apresentam a mesma densidade de probabilidade (isto é, são equivalentes de distribuição (HECKERMAN, 1997)). Isto significa que, caso duas redes possuam exatamente a mesma distribuição, não há condições de se distinguir entre as duas na ausência de conhecimento a priori. Ou seja, em muitas situações, será impossível para um algoritmo qualquer de

inferência de redes bayesianas recuperar exatamente a mesma rede que gerou os dados, mesmo que a amostragem seja infinita, pois há outros modelos que representam os mesmos dados com a mesma parcimônia e a mesma eficiência, sendo, portanto, totalmente equivalentes em termos de complexidade e distribuição.

Mais uma vez, convém considerar que este não é necessariamente o caso aqui. É possível que o algoritmo esteja simplesmente selecionando uma rede ruim. Esta justificativa será abordada na próxima seção.

A.2.4 O K2 como Algoritmo de Maximização

Como comentado anteriormente, o algoritmo K2 é um algoritmo de busca. Ele tenta encontrar a estrutura de rede que maximiza a verossimilhança para um conjunto de dados. Os experimentos realizados na Seção A.2.3 mostraram que nem sempre a rede encontrada corresponde ao modelo original, muitas vezes porque o número de dados utilizados não é suficientemente representativo. Além disso, o critério de máxima verossimilhança influencia o resultado de forma a encontrar modelos menos parcimoniosos. Mas e quanto à eficiência do algoritmo em si? Será que o K2 encontra sempre a rede com verossimilhança máxima dentre todas as possíveis ou ele converge para um máximo local? Em outras palavras, o fato das redes encontradas não terem sido exatamente as procuradas é resultado apenas da falta de representatividade dos dados ou a eficiência do algoritmo K2 também influencia no resultado?

O objetivo desta seção é avaliar o potencial do algoritmo K2 como algoritmo de maximização. Para isso, serão comparadas as verossimilhanças das redes originais com as das redes encontradas. Se a rede encontrada possui uma verossimilhança maior que a da rede original significa que o algoritmo está fazendo o seu papel em maximizar o critério de qualidade. Caso contrário, o algoritmo não está fazendo a busca de maneira adequada, e a sua ineficiência tem uma parcela significativa de responsabilidade nos resultados encontrados.

Quando a Verossimilhança Não Corresponde à Verdade:

Quando a distribuição dos dados observados não corresponde exatamente à densidade de probabilidade do modelo original, é possível que exista uma outra estrutura de

rede bayesiana capaz de representar os dados com uma maior verossimilhança. Neste caso, o compromisso do algoritmo de busca é de encontrar esta outra rede e não a rede original que gerou os dados. Utilizando os mesmos modelos da Seção A.2.3, foram avaliadas as verossimilhanças das redes originais e das redes encontradas quando a representatividade dos dados não é máxima.

Para o experimento 1 da Seção A.2.3, comparamos a verossimilhança da rede encontrada com a da rede original quando o número de amostras é 150 e 250, valores em que as duas redes diferem e a representatividade dos dados é baixa. A Tabela 6.2 mostra os resultados médios obtidos em 20 diferentes amostragens para cada situação. Os valores da tabela são negativos porque a verossimilhança é medida em logaritmo.

Tabela 6.2 Desempenho médio do algoritmo K2 para o problema do experimento 1 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

Nº de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a original (%)**
150	−304,5946	−302,8409	100%
250	−498,2075	−499,3341	90%

Para o problema do experimento 2, foram utilizadas amostragens com 200 e 1000 amostras. Os resultados médios obtidos em 20 amostragens diferentes são mostrados na Tabela 6.3.

Tabela 6.3 Desempenho médio do algoritmo K2 para o problema do experimento 2 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

Nº de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a original (%)**
200	−395,3814	−397,2511	30%
1000	−1.960,4372	−1.957,1221	20%

Para o experimento 3, foram testadas situações com 500 e 1000 amostras. A Tabela 6.4 apresenta os resultados médios.

Tabela 6.4 Desempenho médio do algoritmo K2 para o problema do experimento 3 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

Nº de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a original (%)**
500	−2,2728	−2,2645	100%
1000	−4,5151	−4,5120	100%

Os resultados desta análise são um pouco contraditórios. Para os experimentos 1 e 3, o algoritmo K2 se comportou extremamente bem, encontrando em quase todas as situações uma rede que maximiza a verossimilhança. No experimento 2, no entanto, o desempenho do algoritmo foi bastante ineficiente. A rede original possui quase sempre uma verossimilhança maior que a da rede encontrada. Isso significa que o algoritmo K2 deveria ter sido capaz de recuperar a rede original ou então alguma outra com maior verossimilhança.

Comecemos então analisando o experimento 2. Como dito na Seção A.2.1, o algoritmo K2 é um algoritmo guloso. Uma vez seguindo em uma direção, ele não poderá voltar atrás, convergindo assim para um ótimo local. É possível que, para um dado problema, a introdução de um determinado arco *a* seja melhor em termos de qualidade do que a de qualquer outro arco, mas que dois outros arcos *b* e *c* em conjunto e na ausência de *a* produzam uma estrutura ainda melhor. A questão é que o algoritmo decidirá inicialmente pelo arco *a*, sendo então incapaz de encontrar a melhor estrutura, isto é, aquela que contém *b* e *c*.

Nos outros experimentos isto não aconteceu. O algoritmo encontrou uma solução melhor que a original (embora não saibamos se existe uma outra solução melhor que a encontrada), indicando que a sua busca foi eficiente. Imagina-se, pois, que as superfícies de busca no espaço de estruturas seja menos "acidentado" para estes problemas. Se elas realmente possuírem menos ótimos locais que a superfície de busca do experimento 2, torna-se mais fácil para um algoritmo guloso encontrar a melhor solução.

Através dos testes realizados, não é possível generalizar a conclusão de que o algoritmo é uma técnica boa ou ruim de maximização; conclui-se apenas que ele não é

ótimo. É necessário avaliar o desempenho de outros algoritmos junto ao problema do experimento 2 para realizar uma análise comparativa.

Quando a Verossimilhança é a Verdade:

Quando o número de amostras é suficientemente grande, pelo menos para os problemas simples analisados na Seção A.2.3, é aceitável esperar que não exista outra rede a não ser a original (ou então a sua equivalente de distribuição) que explique melhor os dados observados, isto é, que a verossimilhança é uma medida da verdade. Neste experimento, tentaremos avaliar se em situações desse tipo a rede encontrada pelo K2, quando difere da rede original, é uma equivalente de distribuição. Isto significa dizer que o algoritmo foi competente o suficiente para encontrar a melhor solução (ótimo global), mesmo que a rede não seja exatamente a esperada.

O primeiro teste foi realizado para a rede do experimento 2, na situação em que o número de amostras é 50.000. Espera-se que esse número de amostras seja suficientemente grande para representar fielmente o modelo verdadeiro. O segundo teste foi feito com a rede do experimento 3, também para 50.000 amostras, quando o algoritmo encontra a mesma rede da Figura 6.5(b). Os resultados obtidos são mostrados na Tabela 6.5.

Tabela 6.5 Verossimilhança do modelo original e da rede encontrada pelo algoritmo K2 para os experimentos 2 e 3 com 50.000 amostras.

Experimento	**Verossimilhança do modelo original**	**Verossimilhança da rede encontrada**
2	$-9{,}5158\times10^{4}$	$-9{,}8990\times10^{4}$
3	$-2{,}2267\times10^{5}$	$-2{,}2267\times10^{5}$

No primeiro teste, a verossimilhança da rede obtida (Figura 6.3(b)) é menor do que o da rede original. Isto significa que o algoritmo não teve um bom desempenho, pois as redes não são equivalentes de distribuição. No segundo teste, entretanto, a rede encontrada (Figura 6.5(b)) e a rede original, embora diferentes, possuem exatamente a mesma verossimilhança, ou seja, são equivalentes de distribuição. Se a distribuição dos dados for realmente suficientemente representativa, o algoritmo foi capaz de encontrar o ótimo global.

A.2.5 Discussão

Os métodos de inferência de redes bayesianas são realmente úteis como ferramenta de descoberta das relações causais entre variáveis e de modelagem de distribuição em problemas complexos de mundo real? Referimo-nos mais especificamente a problemas em que o número de variáveis tende a ser grande e a quantidade de amostras é bastante limitada. As redes têm utilidade prática para este tipo de situação?

Embora as análises realizadas aqui sejam insuficientes para responder de forma conclusiva a estas perguntas, baseado nos resultados obtidos é possível arriscar um palpite coerente.

Foi visto que o algoritmo K2 depende de uma quantidade de amostras excessivamente grande – considerando as restrições impostas pelos problemas em foco – para chegar a uma rede que explique perfeitamente os dados (experimentos 1 e 3) e que em algumas situações, nem com um número infinito de amostras é possível recuperar a densidade de probabilidade original (experimento 2) – este último caso deve ser considerado à parte, já que o resultado está relacionado a uma limitação específica do algoritmo que talvez possa ser atenuada com o uso de heurísticas mais eficientes. Conforme discutido na Seção A.2.3, uma rede com apenas 3 variáveis precisa de 1000 amostras para compor um conjunto de dados representativo. Segundo o princípio de maldição da dimensionalidade, uma rede com mais variáveis deve ter o seu conjunto de dados acrescido exponencialmente para que esta representatividade se mantenha. Contudo, na prática, o princípio não se confirmou. Para o experimento 3, envolvendo uma rede com 6 variáveis, com o mesmo número de amostras foi possível encontrar uma rede equivalente à original. Talvez o problema não seja tão crítico assim. Parece que a natureza do modelo é a grande determinante neste caso. A questão é que, se todas as relações causais são bastante intensas, isto é, suas conseqüências são observadas com grande probabilidade, um número relativamente pequeno de amostras é suficiente para compor uma amostragem representativa. Mas se nestes mesmos termos uma das conexões é relativamente fraca, o conjunto amostral deve ser consideravelmente maior para incluir também os eventos menos prováveis de forma significativa. Ora, geralmente não é de estrita relevância ter acesso a esses pormenores, dado que um modelo aproximado contendo apenas as relações causais mais intensas seguramente possuirá robustez suficiente para explicar e generalizar

a maioria dos fenômenos. É, portanto, de fundamental importância que as redes geradas revelem as conexões mais intensas e, para isso, não é necessário um conjunto amostral de tamanho expressivo.

Existe um outro ponto que merece destaque, e se refere às redes equivalentes de distribuição. A análise da Seção A.2.4 mostrou que, em algumas situações, existem redes bayesianas com estruturas diferentes, mas que possuem exatamente a mesma densidade de probabilidade. Como argumentado em HECKERMAN (1997), nesses casos é impossível para qualquer algoritmo fazer a distinção entre os modelos baseando-se apenas nos dados. Isso leva então a um questionamento: o quão diferente podem ser duas redes equivalentes de distribuição e com que freqüência essa particularidade pode ocorrer? Primeiramente, se duas redes equivalentes de distribuição podem apresentar estruturas completamente diferentes, a escolha arbitrária pelo modelo errado pode trazer conseqüências desastrosas quando se está interessado nas relações causais, e não na distribuição em si. Esta, no entanto, não foi a situação observada nos experimentos. Segundo, se a ocorrência de redes equivalentes é freqüente, passa-se a não ter confiança alguma nos resultados encontrados, a não ser que a primeira afirmação esteja errada. Esta é uma questão especial que deve ser investigada com cautela.

Falta comentar sobre o desempenho da abordagem proposta. Os testes mostraram que o algoritmo K2, utilizando como critério de qualidade a máxima verossimilhança, deixou a desejar em várias circunstâncias. Em particular, os experimentos realizados na Seção A.2.4, deixaram claro que o algoritmo converge para ótimos locais com uma certa freqüência, sendo esta uma das razões pelas quais a estrutura original dos modelos não é recuperada. Além disso, foi visto que o critério de máxima verossimilhança tende a valorizar redes mais complexas, o que leva a conexões não existentes na rede original e reduz a aplicabilidade prática dos modelos gerados.

Voltemos então à pergunta inicial. A abordagem empregada para síntese de redes bayesianas pode ajudar a resolver problemas complexos? A conclusão final dos experimentos, embora ainda carente de embasamento em investigações mais profundas, é que sim. Com o uso de uma abordagem mais sofisticada, isto é, com heurísticas de busca mais eficientes e critérios de seleção de modelos mais consistentes, a tarefa de síntese de redes bayesianas sem conhecimento a priori pode ajudar a encontrar as relações mais

intensas entre as variáveis, mesmo na ausência de um conjunto de amostras muito representativo, gerando por sua vez modelos que podem ajudar a entender os eventos associados a problemas de mundo real.