Redes de regulação gênica

Elier Broche Cristo

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Doutorado em Estatística
Orientador: Prof. Dr. Eduardo Jordão Neves

Durante o desenvolvimento deste trabalho, o autor recebeu
auxílio financeiro da FAPESP (processo: 04/00582-0) e CAPES.

São Paulo, janeiro de 2009

A mis padres,

Inés y Juan Antonio,

para quienes mis estudios siempre fueron un sueño.

Agradecimentos

AGRADEÇO A MEUS PAIS, Inés e Juan Antonio, por meus estudos sempre terem sido um sonho para eles e por terem me dado todo o apoio necessário para realizá-los. À minha noiva Débora, pela companhia e apoio constante. Aos meus irmãos Osnel e Elayny, pela confiança e ajuda. Aos meus sogros, Maria e Eduardo, pela motivação. A meu orientador, Eduardo Jordão Neves, pela sua orientação na minha formação. Ao David, pela ajuda no BIOINFO, assim como no lado pessoal e no apoio final. Aos demais colegas do grupo de trabalho, Roberto Hirata, Gustavo Esteves, Ana e Lucas. A todos, pela ajuda nestes cinco anos, correções do português, orientações, explicações, etc.

Agradeço também à FAPESP e à CAPES, pelo apoio financeiro e técnico. Ao BIOINFO/USP, ao Instituto de Matemática e Estatística e à Universidade de São Paulo. A todas estas instituições, agradeço a oportunidade de dar continuidade à minha formação. Considero extraordinária a existência de tais possibilidades no Brasil para estudantes de outros países.

Elier Broche Cristo

São Paulo, SP
5 de janeiro de 2009

Resumo

ESTE TRABALHO APRESENTA UM ESTUDO sobre modelagem de vias de sinalização. Foram empregados, em parceria com o Instituto Ludwig, tanto dados simulados quanto reais em vias de grande interesse em pesquisas associadas ao câncer. São consideradas duas possíveis abordagens: discreta (redes booleanas), contínua (equações diferenciais). Os dados reais usados referem-se a resultados de experimentos com *microarrays* (expressão gênica), onde um conjunto grande de genes é monitorado ao longo do tempo em diferentes condições biológicas. Os resultados obtidos demonstram a adequação dos modelos utilizados na descrição de forma simples e relativamente clara o comportamento de vias de sinalização com enfoque qualitativo. Dessa forma, geramos ferramentas que permitem a estimação de parâmetros em dados reais e simulados que explica a interação entre os genes e ajuda a entender as diferenças existentes entre duas condições biológicas.

Palavras-chave: vias de sinalização, modelagem com equações diferenciais, expressão gênica, *microarray*, câncer.

Abstract

THIS IS A STUDY on network signaling modeling. In partnership with Ludwing Institute have been used both simulated and real data in networks that have great interest in studies on cancer. Two possibilities are considerated: discret (Boolean network) and continuous (differential equations). The real data used refers to results of experiences with microarrays (genic expression), where a great group of genes is monitored over time in different biological conditions. The results have been shown to be appropriate in the description of a simple and clear in the conduct of signaling networks focusing on qualitative analysis. In this way, we made tools that allow the estimation of the parameters in real and simulated data which explains the interaction between the genes and helps to understand the differences between two biological conditions.

Keywords: signaling networks, differential equations modeling, genic expression, microarray, cancer.

Sumário

Lista de Figuras

Lista de Tabelas

Apresentação

A PRESENTE VERSÃO DESTE DOCUMENTO é própria para impressão. Uma versão eletrônica adequada para ser lida em um computador, contendo apontadores para as páginas citadas e para referências cruzadas, pode ser encontrada na URL citada na ficha técnica, última seção desta apresentação. Foi evitado ao máximo, em todo o texto, a referência a endereços na internet, devido às recorrentes mudanças que estes sofrem.

Espera-se que este documento seja útil para o aprendizado de alguns conceitos que foram dominados durante o doutorado e que desperte o interesse por uma continuação da pesquisa nesta área, que ainda tem tanto por fazer.

Este trabalho iniciou-se como parte das atividades do projeto CAGE e do BIOINFO/USP. Além de acesso às instalações computacionais e *softwares* do IME-USP, temos também acesso aos recursos adicionais do BIOINFO/USP. Este núcleo, possui dois laboratórios com cerca de 30 computadores. Além disso, temos um servidor e dois computadores dedicados ao processamento no Maiges/Numec. Temos acesso a vários conjuntos de dados gerados no Instituto Ludwig através da colaboração com o Prof. Dr. Luiz Fernando Reis. Também contamos com ampla disponibilidade dos importantes *softwares* necessários para a análise que realizamos, em particular: R, Xppaut e Matlab.

Ficha Técnica

Seguem alguns dados referentes ao doutorado, às máquinas usadas para os experimentos e ao conteúdo deste documento.

☞ Doutorado:

Início: 1º de janeiro de 2004.

Exame de qualificação: março de 2008.

Formação anterior: Mestre em Ciências, Área de Concentração em Estatística—IME/USP.

☞ Conteúdo:

Assunto: Modelagem de vias de sinalização.

Palavras-chave: vias de sinalização, modelagem com equações diferenciais, expressão gênica, *microarray*, câncer.

Versão eletrônica: `http://www.vision.ime.usp.br/~elier/doutorado/tese.pdf`

1
Introdução

Este trabalho apresenta um estudo sobre modelagem de redes de regulação gênica. Numa primeira aproximação, a evolução de uma célula é determinada pela evolução dos níveis de expressão de seus genes [Weng *et al.*, 1999]. Embora, sem dúvida, desconsidere fenômenos importantes da dinâmica celular, esta visão torna o problema tratável tanto do ponto de vista experimental quanto do ponto de vista teórico. O monitoramento dos níveis de expressão de milhares de genes simultaneamente é possível através do emprego da tecnologia de *Microarray* [Nat, 1999; Schena, 2002]. A análise matemática e estatística e a posterior interpretação dos dados gerados por esta tecnologia têm sido objeto de muitas pesquisas [Laub *et al.*, 2000; Rocke & Durbin, 2001]. Redes de regulação gênicas podem ser analisadas em diversos níveis de abstração e por diversas abordagens matemáticas como, por exemplo: *Boolean Networks*, HMM (*Hidden Markov Models*), *Bayesian Networks*, *Stochastic Models*, *Differential Equations*, MCMC (*Markov Chain Monte Carlo*), *Linear Models* e *Neural Networks*. Particularmente interessante, nos parece a abordagem envolvendo *Piecewise Linear Differential Equations* [Gouzé & Sari; Jong, 2002; Jong *et al.*, 2003; Schneider *et al.*, 2003], um meio termo entre uma abordagem discreta e uma abordagem contínua, que está bem adaptada, ao nível de informação quantitativa usualmente disponível sobre redes de regulação gênica.

Dispomos atualmente de um protótipo de ambiente matemático/computacional [Esteves, 2007], adequado para analisar dados reais, desenvolvido no nosso grupo de pesquisa e do qual o aluno participou. Neste ambiente computacional, podemos testar e comparar diversas das abordagens citadas. Por exemplo, trabalhamos em problemas envolvendo redes de regulação gênica em pesquisa sobre o câncer, em colaboração com o Instituto Ludwig. Um desafio importante é o de desenvolver métodos capazes de comparar objetivamente, com adequado rigor estatístico, as inferências das diferentes abordagens em diferentes situações concretas. Nesse propósito, um dos objetivos deste trabalho foi estudar sistemas dinâmicos de expressões gênicas, como continuação natural do trabalho de pesquisa de mestrado, do aluno, que contou com apoio financeiro da FAPESP (Processo: 01/07082-5), voltado ao estudo de métodos estatísticos na análise de experimentos de *microarray* [Cristo, 2003]. Foram considerados problemas envolvendo redes de regulação gênicas em diversos contextos reais, em pesquisas sobre o câncer. Este projeto faz parte de uma produtiva colaboração com o Instituto Ludwig, com importante participação do aluno [Gomes *et al.*, 2005, 2003; Meireles *et al.*, 2004; Stolf *et al.*, 2006]. O presente trabalho iniciou-se associado ao projeto temático

interdisciplinar IME-USP/IQ-USP FAPESP *CAGE*, *Cooperation for the Analysis of Gene Expression*, ao projeto temático IME-USP *Fenômenos Críticos em Processos Evolutivos e Sistemas em Equilíbrio*.

O aluno, em sua dissertação de mestrado, analisou diversos problemas associados à expressão gênica, abordando questões que envolvem desde importantes etapas iniciais de pré-processamento (análise de imagem e normalização), como questões de agrupamento e classificação. No que refere-se a classificação, a pesquisa foi encaminhada a encontrar marcadores moleculares baseados em "gene cliques", pequenos grupos de genes capazes de distinguir amostras em condições biológicas diferentes, satisfazendo determinados critérios de qualidade. Foram identificados diversos cliques de genes que exibem, nas amostras, padrões de expressão suficientemente diferentes, nas distintas condições biológicas de interesse (tumor benigno ou metastático, por exemplo) para sugerir que possam levar à construção de bons classificadores (marcadores moleculares).

Uma vez identificados os classificadores, o problema que segue naturalmente é o de identificar quais são as interações entre esses genes, qual é a rede de interação entre eles, cuja alteração dá lugar à diferença de padrões observada. Esta é uma área de intensa atividade de pesquisa atual. Este problema pode ser abordado de maneiras bastante diferentes e em diversas escalas [Jong, 2002; Smolen *et al.*, 2000]. Do ponto de vista matemático, há diversos desafios. Como em diversas outras situações reais, o grafo associado a redes de interação gênica (genes correspondendo aos nós e elos indicando a existência de interação entre o par de genes correspondentes) parece ter as propriedades de grafos aleatórios bastante interessantes [Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002]. Há evidências, por exemplo, de que este grafo pode ter características de grafos livres-de-escala, com genes de alta conectividade correspondendo a genes supressores de tumores ou oncogenes. A própria estrutura de conectividade global desse grafo, apresenta uma evolução dinâmica que parece ser muito importante na evolução de sistemas biológicos. Sistemas dinâmicos definidos sobre este tipo de grafos tem diversas propriedades, como existência de transição de fase dinâmica, que ainda não estão bem compreendidas matematicamente e cuja relevância no contexto biológico ainda está para ser estabelecida.

Uma via de sinalização pode ser resumida a um conjunto de ações específicas (reações químicas) na célula, onde o sinal é passado entre os componentes bioquímicos (exemplo: proteínas e genes) que a formam. Estas vias realizam o controle das funções da célula, tais como divisão celular e a morte celular programada. O fluxo de informação em uma via de sinalização ocorre de tal forma que pode ser capaz de realizar complexas tarefas de procesamento de sinais, fazendo parte do complexo controle dos diversos processos que têm lugar na célula. As estratégias de controle podem ser muito simples ou sofisticadas, e elas têm a finalidade de evitar falhas nos módulos que formam as vias de sinalização, pois caso isto ocorra fica comprometido o funcionamento adequado das funções fisiológicas na célula, podendo gerar doenças como o câncer. Também podemos pensar em uma via de sinalização como um conjunto de módulos funcionais conectado entre si pelos próprios fluxos de informação, estes módulos podem ser identificados pela conexão que existe entre seus componentes e/ou funções específicas.

As vias são definidas pelos fluxos de informação entre seus componentes bioquímicos e pode ser dividida em pequenos módulos com funções específicas. Estes são dois dos fatores que a fazem de grande interesse para estudo. Dentro de uma via de sinalização podemos estudar independentemente cada um dos módulos

que a formam e a via como um todo. Quando temos dados reais de série temporal para duas condições biológicas (exemplo: Tumor, Normal) podemos estudar as diferenças na sinalização entre as duas condições biológicas e inclusive tentar conjecturar qual é o problema na condição patológica do câncer. Muitos dos problemas em aberto atuais, associados ao câncer, por exemplo, provavelmente têm uma explicação ao nível genético. Conhecendo com mais detalhes os genes envolvidos e as interações entre eles pode ser estudado quais são as diferenças entre tecidos normais e tecidos tumorais. Tendo estas diferenças pode ser um começo para tentar evitar diagnósticos tardios, assim como pensar em possíveis tratamentos. Alguns exemplos de vias de sinalização estudadas neste trabalho são: EGF, AKT, Wnt, Apoptosis, Cellcycle, Mapk Erk e Jak Stat. Estas e outras vias de sinalização podem ser encontradas no banco de vias de sinalização STKE (`http://stke.sciencemag.org/cgi/collection/canonical_pathways`).

A princípio existem duas grandes abordagens que podem ser feitas na modelagem de redes de regulação gênica: discreta e contínua. No começo deste trabalho iniciamos a pesquisa centrando a nossa atenção na modelagem discreta e atualizações simultâneas. Umas das maiores dificuldades quando consideramos densidades discretas é a determinação do nível expressão que define, abaixo do qual, ou acima do qual, o estado de um determinado gene muda. Note que a determinação destes pontos para os diferentes genes não é igual nem trivial. Quando trabalhamos com dados reais, o número de genes medidos e instantes de tempo é limitado, o que faz a estimação destes pontos mais difícil. A atualiação síncrona é mais simples, mas ao mesmo tempo não parece a mais adequada. Se consideramos que para genes diferentes as escalas de tempo, nas quais acontecem as mudanças podem ser muito diferentes, então é esta mais uma razão para pensarmos em outro tipo de modelo. Parece fazer mais sentido uma modelagem, na qual tanto o tempo como os níveis de expressão sejam considerados como contínuos, inclusive com atualizações assínconas. São estas algumas das razões pelas quais acreditamos que uma modelagem por equações diferenciais seja muito adequada para redes de regulação, onde as taxas sejam definidas pelas forças de interação existentes entre os genes.

O objetivo principal deste trabalho é desenvolver modelos que descrevam de forma simples e clara o comportamento de vias de sinalização com enfoque qualitativo. Desta forma, captar com precisão e embasamento biológico as regras de interação existentes entre os genes e verificar a existência de diferenças significativas entre duas condições biológicas diferentes. Em resultados obtidos neste trabalho, com dados reais e simulados, verificamos que a modelagem contínua se adéqüa muito bem e identifica as forças de regulação existentes entre os genes na via ao longo do tempo, além de captar também as diferenças existentes entre condições biológicas diferentes para uma determinada via. Um exemplo disto são os resultados obtidos para as duas condições biológicas no conjunto de dados com células imortalizadas. Neste conjunto de dados (vide Section 6.0.1) temos duas séries temporais, cada uma referente a uma condição biológica: tratamento (Sulindac) e controle (DMSO). Verificamos a consistência do método entre as diferentes estimativas, assim como as possíveis diferenças genéticas das células imortalizadas nas duas condições biológicas em estudo, desde o ponto de vista de forças de interação gênica. Já nos dados simulados podemos verificar a eficiência do método e o poder de estimação, tanto no baixo nível de erro como na consistência entre as estimativas obtidas.

A principal dificuldade nos experimentos de *microarrays* é a falta de informação devido ao fato de que apenas estão envolvidos no experimento biológico alguns dos genes presentes na rede. Ou seja, não temos

as medições dos níveis de expressão para todos os genes presentes na via. A estimação dos parâmetros presentes no modelo é realizada com apenas as séries temporais dos genes medidos, dados reais. Além dos níveis de expressão também temos a informação a priori que têm-se da via, os genes que a formam e as interações entre eles. Outra grande dificuldade existente é o grande número de parâmetros envolvidos no modelo, $N_p = 2N_z + 2N_{ne} + N_{nm}$, em que N_p corresponde ao número total de parâmetros estimados, N_z é o número e interações (zetas), N_{ne} define o número de genes não entrada e N_{nm} o número de não medidos. Ou seja, cada interação entre dois genes gera um α e um β, cada componente não entrada gera um a e um b e cada gene não medido gera um x_0 (nível de expressão inicial), sendo estes os parâmetros considerados na modelagem. Não serão considerados no modelo genes que não tenham pelo menos uma interação com algum outro gene.

A estimação é feita da seguinte forma: primeiramente são obtidos vários conjuntos de estimativas dos parâmetros envolvidos no modelo e, posteriormente pode ser avaliada a qualidade do ajuste para cada um dos conjuntos de estimativas como para todas as estimativas de um determinado parâmetro. O resultado do algoritmo de estimação por SA é uma estimativa para cada parâmetro, mas quando rodamos este algoritmo N_s (número de simulações, estimações) vezes para a mesma via, obtemos N_s conjuntos de estimativas, sendo que para cada uma destas simulações podemos avaliar a qualidade do ajuste. O resultado final do processo de estimação é uma estimativa para cada um dos parâmetros, sendo possível determinar este conjunto de estimativas finais de duas formas possíveis:

① aquele conjunto de estimativas com a melhor qualidade de ajuste, escolhido entre os N_s conjuntos de estimativas;

② criar um conjunto de estimativas novo a partir dos N_s conjuntos obtidos, em que a estimativa para um determinado parâmetro pode ser considerada como uma medida resumo entre as N_s estimativas para esse parâmetro.

A medida de qualidade para cada conjunto pode ser considerada como o somatório do erro quadrático, $\sum_{i=1}^{N_m}\sum_{j=1}^{N_t}(x_{ij}^e - x_{ij}^r)^2$, ou o valor máximo do erro quadrático, $\max_{\{i\in[1,N_m],\ j\in[1,N_t]\}}(x_{ij}^e - x_{ij}^r)^2$. N_t é o número de tempo da série temporal de interesse, N_m corresponde ao número de genes medidos no experimento, x^e e x^r denotam os níveis de expressão gênica estimadas e reais, respectivamente.

Podemos separar os genes em entradas (aqueles em que nenhum dos outros influenciam) e não entrada (pelo menos um gene o influencia). Também podemos dividir os genes em outros dois grupos, medidos (aqueles para os quais temos os níveis de expresão reais) e não medidos (aqueles que não foram incluídos no experimento). Se cruzamos estas duas divisões podemos gerar quatro grupos: os entrada medidos, os entrada não medidos, os não entrada medidos e os não entrada não medidos. Desta forma, não serão considerados os genes que sejam entrada e que não sejam medidos pelo fato de não aportar informação útil, os outros três tipos sim serão considerados. Dos entrada medidos não conseguiremos gerar as equações diferenciais, mas sim temos a evolução dos níveis de expressão reais ao longo do tempo, porém não poderemos usar esta informação para avaliar a qualidade do ajuste pois pelo fato de não termos as equações diferenciais para eles não será possível gerar os níveis de expressão simulado ao longo do tempo. Por outro lado os não entrada sim terão equação diferencial e consequentemente níveis de expressão simulado, mas para a avaliação da

qualidade do ajuste apenas poderão ser considerados os não entrada medidos pois para os não entrada não medidos não temos os níveis de expressão reais.

1.1 O problema biológico

Um dos novos desafios em Biologia teórica e molecular consiste em encontrar possíveis representações para numerosos componentes moleculares, interagindo para o controle e fisionomia, que envolve um organismo completo. As redes de regulação gênica são um caso particular deste problema geral. Para entender a natureza das funções celulares, primeiro é necessário estudar o comportamento dos genes. Tanto as expressões como as atividades dos genes não são isoladas nem independentes. Esta é a razão de se estudar o todo, como uma rede de elementos que estão conectados e interagindo entre si. Muitos modelos matemáticos e computacionais têm sido desenvolvidos com a finalidade de modelar as interações gênicas, anteriormente explicadas. Algumas questões de interesse são:

① Conhecimento da estrutura geral da rede;

② Presença ou não de interações cruzadas;

③ Consideração ou não dos circuitos de *feedback*;

④ Idéia da rede como um conjunto de sub-redes conectadas.

2
Revisão Bibliográfica

SEGUE UM CONJUNTO DE IDÉIAS existentes na bibliografia e que consideramos relevantes na modelagem de redes de regulação gênica [Laub *et al.*, 2000; Schneider *et al.*, 2003].

2.1 Boolean Networks

A modelagem das interações entre os genes, empregando *Boolean Networks* é realizada através de um diagrama de estados. Neste processo, o perfil de expressão dos genes pode ser considerado como *on* ou *off*, definindo os estados dos genes. Desta forma, pode ser modelada uma *Boolean Network*, na qual os genes são os nós e as interações entre eles são as ligações entre os nós. Depois de terem sido criadas as *Boolean Networks*, podem ser inferidas as regras da rede, usando para isto um determinado algoritmo de aprendizado.

Uma aproximação comumente usada, para modelar rede de expressão gênica complexa (elementos e interações) é mediante equações booleanas [Thieffry & Romero, 1999]. Neste contexto, circuitos de *feedback* (seqüências circulares de interações) têm-se mostrado como muito adequados, visto que circuitos positivos são habilitados para gerar *multistationarity*. Circuitos negativos podem gerar comportamento oscilatório. Então, podemos pensar em uma rede de regulação gênica representada através de uma rede booleana, conectada completamente, na qual cada elemento interage com todos os elementos, inclusive com ele mesmo. No desenho da rede é introduzida certa flexibilidade pelo uso dos parâmetros booleanos, associado com cada interação ou grupo de interações, afetando um determinado elemento. Dentro deste formalismo, um circuito *feedback* pode ser utilizado na geração de comportamento dinâmico típico (*multistationarity* ou oscilações), apenas para valores apropriados, de alguns dos parâmetros lógicos. No caso em que as condições anteriormente colocadas são satisfeitas, podemos afirmar que o circuito é "funcional" [Thieffry & Romero, 1999]. Usando esta metodologia, pode ser encontrada a proporção de combinações de valores de parâmetros que fazem um circuito funcional geometricamente decrescente com o comprimento do circuito. Do ponto de vista biológico, podemos pensar que redes de regulação poderiam ser divididas em pequenos circuitos *feedback* ou módulos regulatórios que sejam relativamente independentes.

2.2 Bayesian Networks

As *Bayesian Networks* são um caso especial de uma classe mais geral, conhecida como modelos gráficos, na qual os nós representam variáveis aleatórias e a falta de conexão entre os nós indica a suposição de independência condicional. Este tipo de rede está formada por duas partes: um DAG (*Directed Acyclic Graph*) e um conjunto de parâmetros. O DAG é um conjunto de nós que representam variáveis aleatórias e arcos que representam as relações entre as variáveis aleatórias. O conjunto de parâmetros para cada nó no DAG, especifica uma distribuição conjunta das variáveis aleatórias. A distribuição conjunta é decomposta no gráfico usando considerações de independência condicional.

Nas *Bayesian Networks*, os arcos direcionados representam relações causa-efeito. As evidências podem ser atribuídas a qualquer um dos nós de um arco, estas redes são geralmente usadas em dois tipos de casos:

① de causa conhecida para efeito desconhecido;

② de efeito conhecido para causa desconhecida.

Estes tipos de redes, também fazem uma redução do espaço paramétrico, pela consideração de probabilidades condicionais.

As técnicas de agrupamento, as quais têm uma implementação relativamente simples, e as *Boolean Networks*, as quais são ainda mais intuitivas, não são adequadas para modelar redes de regulação gênica no tempo [Poornachandran & Chidambaram, 2003]. *Dynamic Bayesian Networks* modelam as dependências condicionais e consideram as mudanças nos perfis de expressão dos genes no tempo [Poornachandran & Chidambaram, 2003].

2.3 Circuito de *feedback* e seus papéis

Os circuitos de *feedback* podem ser definidos de forma simples como seqüências circulares de interações. Algumas propriedades deste tipo de circuito são o sinal (+ ou -) e o número de interações. No entanto, pode-se verificar facilmente, dado um determinado circuito, que cada elemento exerce um efeito similar (positivo ou negativo) nele mesmo. Conseqüentemente, qualquer circuito pode ser colocado em uma de duas categorias: "circuito positivo" ou "circuito negativo". Para conhecer o sinal de um circuito, é suficiente contar o número de interações negativas nele. Se o número é par então o circuito é positivo, caso seja ímpar, o circuito é negativo.

O que faz estes conceitos importantes, é que especificações biológicas e propriedades dinâmicas podem ser associadas com um destes dois tipos de circuitos. Os circuitos positivos estão associados à geração de *multistationarity*. Por outro lado, os circuitos negativos estão associados à sustentação de oscilações (variações periódicas dos valores nas variáveis correspondentes) ou homeostase (confinamento das variáveis envolvidas em volta de algum valor central dos possíveis valores).

No contexto dos formalismos lógicos generalizados, um determinado circuito *feedback* gerará os correspondentes comportamentos, somente para valores apropriados de alguns dos parâmetros lógicos. Sempre que acontece o anteriormente explicado, o circuito é chamado de *funcional*.

Com a finalidade de realizar alguma introspecção inicial, na estrutura das redes de regulação biológicas, podem ser empregados critérios de parâmetro lógico, circuitos de *feedback* e a funcionalidade do circuito para caracterizar a estrutura da grande rede booleana. Mais especificamente, poderíamos nos fazer as seguintes perguntas:

① Para uma rede com n elementos, qual é o tamanho do conjunto de combinações consistentes de parâmetros lógicos?

② Que proporção deste conjunto forma um determinado circuito com a propriedade de ser *funcional*?

③ Como fazer para que esta proporção dependa do número de elementos envolvidos no circuito?

Tem-se confirmado que alguns dos requerimentos para que um sistema tenha dinamismo caótico [Toni *et al.*, 1999] são:

① pelo menos um *loop* negativo em um par de elementos;

② um *loop* positivo na estrutura de sistemas de *loop*. Para analisar este dinamismo caótico, pode ser empregado um método *loops* de *feedback*.

2.4 Construção da rede usando informação mútua

O problema de redes de regulação gênica pode ser abordado através de dados de expressão usando, para isto, redes booleanas probabilísticas. O processo de construção da rede pode ser dividido em diferentes etapas [Zhou *et al.*, 2003]:

① Determina-se o número de possíveis conjuntos de genes pais e para cada gene os correspondentes conjuntos de entrada. Isto pode ser feito usando uma técnica de agrupamento baseada em minimização da informação mútua. Neste caso, para resolver o problema de otimização, é empregado *simulated annealing*. Depois destes agrupamentos iniciais de genes, o interesse é nas classes de funções diferentes dos possíveis conjuntos de genes pais, para cada gene *target*.

② Cada função, então, é modelada através de um termo linear e outro não linear. A técnica de MCMC pode ser usada para calcular a ordem do modelo e os parâmetros.

③ Finalmente, o coeficiente de determinação é empregado na determinação da probabilidade de seleção de preditores diferentes para cada gene. Para testar esta aproximação na construção de redes de regulação gênica, poderão ser utilizados diferentes conjuntos de dados que fazem parte deste projeto.

Uma possível abordagem usando informação mútua é a seguinte [Zhou *et al.*, 2003]: para todo par de gene, é realizado o estudo de informação mútua e é escolhido um umbral de informação mútua para criar o agrupamento a partir dos que tem informação mútua maior que um determinado *threshold*. Esta abordagem é baseada nos pares, para os quais temos informação mútua, e assim essencialmente só são analisadas as distribuições marginais dos dados multidimensionais. No caso que empregamos a técnica de agrupamento,

a idéia é minimizar a informação mútua existente nas variáveis entre os grupos e estudar as distribuições de probabilidade conjunta adjacente dos dados. Finalmente, nesta abordagem, o objetivo é a construção das funções da rede para um gene em função dos outros genes a partir da informação obtida do agrupamento. Para estimar as funções de cada gene pode ser usado um modelo com uma componente linear e outra não linear.

Como primeiro passo na construção de uma rede de regulação gênica, a complexidade dos dados de expressão é reduzida por mudanças de *thresholding* nos níveis de transcrição, usando para isto três possíveis valores (-1, 0, 1), os que representam sub-regulado, invariante e sobre-regulado, respectivamente. Quando é usado *microarray*, a intensidade absoluta do sinal varia extensivamente devido ao processo de preparação e impressão dos elementos EST e o processo de preparação e marcação das representações do cDNA dos *pools* de RNA. Este problema, pode ser resolvido por padronização interna. Um primeiro algoritmo calibra os dados, individualmente, para cada lâmina e as expressões são classificadas como (-1, 0, 1), obtendo-se, para isto, a significância estatística. Então uma rede pode ser construída usando a metodologia de agrupamento e predição.

No nosso caso, em que pretendemos trabalhar a partir de dados de expressão gerados com a técnica de *microarray*, os dados são muito limitados para o número de variáveis no sistema. Conseqüentemente, isto é necessário para restringir o número de variáveis para as quais a função está definida. Felizmente, muitas funções só requerem algumas variáveis essenciais.

A motivação para considerar informação mútua é a sua capacidade de medir a dependência geral entre as variáveis aleatórias. A idéia principal desta técnica, é identificar um número relativamente pequeno de pais candidatos para cada gene baseando-se em estatísticas tais como a correlação. Desta forma, resulta um espaço de busca menor, no qual pode ser procurada, de forma mais eficiente, uma "boa" estrutura.

2.5 Entropia

A teoria da informação de Shannon também pode ser usada e muito bem adequada neste contexto de redes de regulação gênica. A entropia da amostra de expressão gênica é uma medida de incerteza da informação contida na amostra. Dado um vetor X e sua distribuição de probabilidades $P(x = x_i)$, $i = 1, 2, \ldots, N_x$, na qual N_x é o número de possíveis valores para X, a entropia é definida como segue:

$$H(x) \triangleq \sum_{i=1}^{N_x} P(X = x_i) \log(P(X = x_i)).$$

2.6 Método de busca de sub-redes ativas

A seguir, é apresentado um método geral de busca [Ideker *et al.*, 2002] na rede, para encontrar sub-redes ativas (conjuntos de genes conectados) para os quais não se espera ter níveis altos, diferencialmente expressos. Depois de ter observado os níveis de expressão em diferentes condições, o interesse é determinar as condições que afetam significativamente os níveis de expressão, isto para cada sub-rede ativa. Com este fim, pode ser implementado um sistema estatístico de *scoring* para determinar a quantidade de mudança

na expressão gênica em uma determinada sub-rede. Este algoritmo de busca é baseado em um processo recursivo (*simulated annealing*). Este processo recursivo pode ser aplicado de modo geral em qualquer problema de otimização ou aprendizado, empregando atualizações sucessivas, as quais podem ser aleatórias ou determinísticas, e o número de iterações é mais um parâmetro do algoritmo. Usando este método, podem ser identificadas as sub-redes de interesse. Pode-se pensar nestes algoritmos em dois possíveis casos:

① uma pequena rede de interações trabalhando com dados de expressão, de condições simples;

② uma grande rede de interações, observada em diversas condições.

2.7 Cálculo do *z-score* básico

Para medir a atividade biológica de uma determinada sub-rede, pode-se começar pela determinação da significância para cada gene como diferencialmente expresso. Para determinar os p-valores, podemos empregar o teste *Mann-Whitney*, p_i representa a significância na mudança de expressão para o i-ésimo gene. Depois de ter determinado os p_i, cada um deles é transformado para um *z-score*, $z_i = \Phi^{-1}(1 - p_i)$, sendo Φ a função de densidade acumulada da distribuição Normal. Desta forma, em dados aleatórios, os p_i são distribuídos segundo uma $U(0, 1)$ e os *z-score* segundo uma $N(0, 1)$. Neste caso, teremos que para um p_i pequeno, corresponde um *z-score* grande. A seguir, pode ser definido *z-score* acumulado:

$$z_A = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} z_i,$$

na qual A corresponde a uma sub-rede de k genes. Note que, depois de termos definido esta medida, será possível fazer comparações entre quaisquer sub-redes, embora elas não tenham o mesmo tamanho (quantidade de genes). A partir deste ponto, poderemos identificar sub-redes ativas biologicamente, aquelas para as quais obtemos valores grandes de z_A.

2.8 Método de busca de sub-redes ativas, caso de muitas condições

O método que tem sido apresentado, pode ser estendido ao caso mais geral, quando temos um grande conjunto de condições. Neste caso, emprega-se uma matriz contendo os p_i, as linhas representando os genes e as colunas as condições, e uma matriz com os *z-score* correspondentes. Dada uma sub-rede A, pode ser empregada a equação definida anteriormente para gerar os *z-score* acumulados $(z_{A1}, z_{A2}, \ldots, z_{Am})$, isto no caso de estarmos estudando m condições. Posteriormente, os *z-score* são ordenados de maior a menor $(z_{A(1)}, \ldots, z_{A(j)}, \ldots, z_{A(m)})$. Segue o cálculo da significância, $r_{A(j)}$ do j-ésimo maior *z-score* usando uma estatística de ordem binomial. A formulação é apresentada a seguir:

Seja $P_z = 1 - \Phi(z_{A(j)})$ a probabilidade de que qualquer condição simples tenha *z-score* sobre $z_{A(j)}$. Então:

$$p_{A(j)} = \sum_{h=j}^{m} \binom{m}{h} (P_z)^h (1 - P_z)^{m-h}.$$

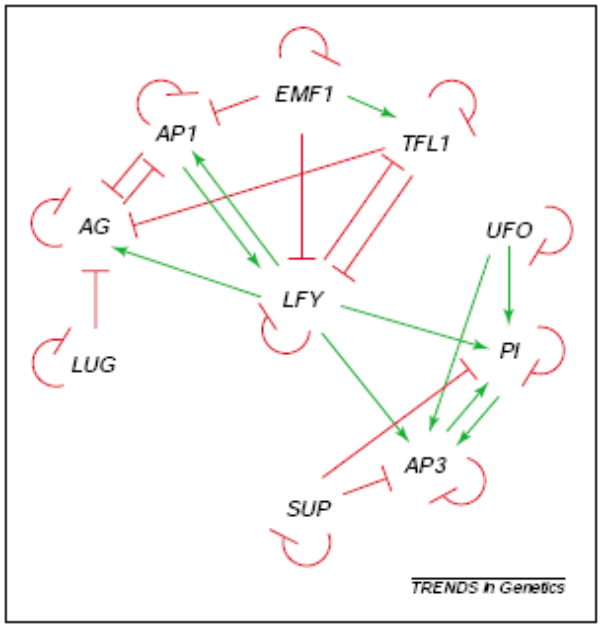

Figura 2.1: *Exemplo de representação gráfica de uma rede [Schneider* et al.*, 2003]. A interação representada por uma seta indica que o gene origem ativa o gene destino, enquanto a interação representada por uma barra indica que o gene origem reprime o gene destino.*

Note que $p_{A(j)}$ é a probabilidade de que pelo menos j das m condições tenham *z-score* sobre $z_{A(j)}$. Pode ser usada a função de densidade acumulada inversa, $r_{A(j)} = \Phi^{-1}(1 - p_{A(j)})$, para converter para a $N(0, 1)$. A seguir define-se:

$$r_A^{\max} = \max_j(r_{A(j)}),$$

o *score* da nova sub-rede.

2.9 Busca de sub-redes com *high-scoring* usando *simulated annealing*

Para o método de busca de sub-redes ativas, foi apresentado como determinar os *scoring* dada uma determinada sub-rede, mas de modo geral também estaremos interessados em encontrar todas as sub-redes, para as quais têm-se os maiores *scoring*. Uma possível aproximação pode ser realizada empregando *simulated annealing* (Kirkpatrick *et. al.*, 1983). Na prática, esta aproximação não garante encontrar o *score* global, no entanto toda sub-rede com *scoring* alto é de grande interesse biológico.

Uma Rede de Regulação Gênica é uma representação gráfica ou numérica das interações entre os genes que formam a mesma, como pode ser observado nas Figuras 2.1 e 2.2, respectivamente.

O coeficiente de co-controle, dado por

$$^{V_m}O^i_j = \frac{\Delta[mRNA_i]/[mRNA_i]}{\Delta[mRNA_j]/[mRNA_j]}, \tag{2.1}$$

caracteriza como as concentrações do $[mRNA_i]$ e do $[mRNA_j]$ mudam segundo uma perturbação da taxa de transcrição V_m de um terceiro $mRNA_m$, sendo que V_m representa a taxa de transcrição do m-ésimo $mRNA$,

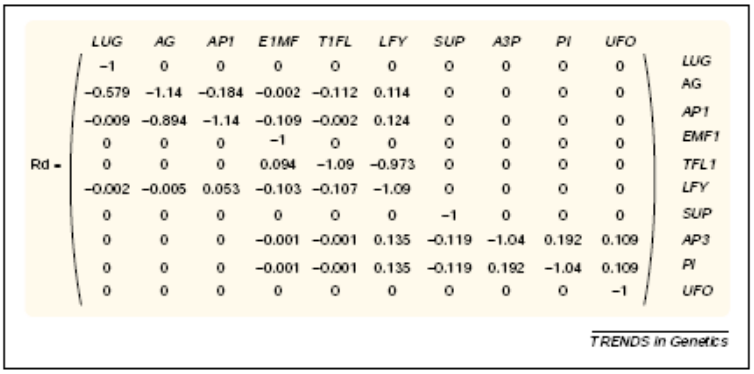

Figura 2.2: *Exemplo de representação numérica de uma rede [Schneider* et al.*, 2003].*

onde *mRNA* são os valores de concentração, i e j são os índices dos genes e $\Delta[mRNA_i]/\Delta[mRNA_j]$ é a razão entre as concentrações dos $[mRNA]$ para os genes i e j respectivamente. A medida ${}^{V_m}O^i_j$ é definida para qualquer i e j, inclusive para $i = j = m$. Também podemos definir $Rd = [{}^jR^j_i]$ que representa a rede gênica pelas interações entre os genes, $Rd = O^{-1}$. Sendo que $O = [{}^jO^i_j]$, na qual ${}^jO^i_j$ é determinado fazendo-se $m = j$ na Equação (2.1).

No caso particular de experimentos com *microarrays* [Cristo, 2003], os níveis de expressão das observações de interesse e de referência são denotadas por $[mRNA]$ e $[mRNA]^0$. Denotamos a razão das intensidades por $FR = [mRNA]/[mRNA]^0$ e a diferença de concentração por $\Delta[mRNA]/[mRNA]^0$,

$$\frac{\Delta mRNA}{mRNA^0} = \frac{mRNA - mRNA^0}{mRNA^0} = FR - 1.$$

Também verifica-se que,

$${}^{V_m}O^i_j \approx \frac{(FR_i - 1)(FR_j + 1)}{(FR_j - 1)(FR_i + 1)},$$

na qual FR_i refere-se ao FR calculado para o i-ésimo gene.

As forças de regulação são boas medidas quantitativas das interações gene-a-gene, e estas podem ser determinadas diretamente a partir de níveis de expressão gênica relativos, como obtidos em experimentos de *microarray*. A habilidade para criar redes gênicas a partir de dados experimentais e usá-los para avaliar sua dinâmica e princípios de desenho, incrementa o conhecimento da função celular.

Uma via de regulação pode ser definida de diferentes formas, segue algumas das possíveis definições: (i) redes metabólicas que representam as transformações químicas entre metabólitos; (ii) redes de proteínas que representam as interações proteína-a-proteína, tais como formação de complexos e modificações pós-traducionais por enzimas de sinalização (conhecidas também como redes de sinalização); (iii) redes gênicas que representam as relações que podem ser estabelecidas entre genes, quando observamos como o nível de expressão de cada um afeta o nível de expressão dos outros. Estas redes podem ser representadas de diferentes formas, no intuito de expressar a maior quantidade de informação da mesma; como por exemplo, as utilizadas nas Figuras 4.1, 2.1, 2.2, 2.3, 2.4 e 2.5.

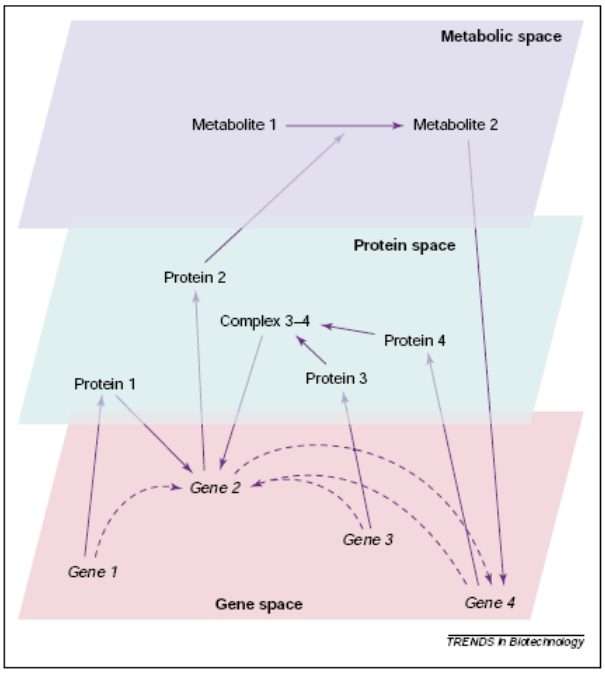

Figura 2.3: *Um exemplo de uma rede bioquímica.*

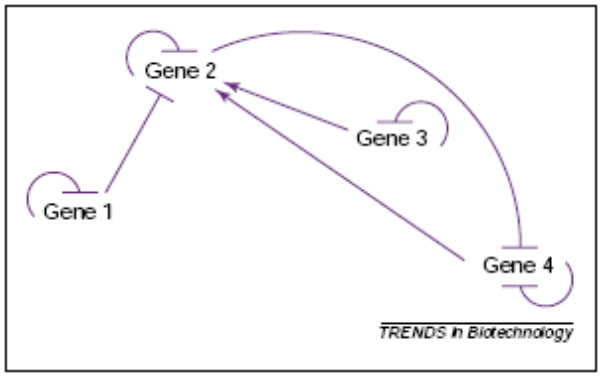

Figura 2.4: *Representação gráfica da rede bioquímica representada na Figura* 2.3.

2.10 Um método quantitativo para engenharia reversa em redes de regulação gênica

Define-se a razão entre os níveis de expressão do estado estimulado e de referência por FR_i,

$$FR_i = \frac{F'}{F} = \frac{[mRNA_i]'}{[mRNA_i]},$$

na qual F' e F são os níveis de intensidade do estado estimulado (por experimento) e de referência respectivamente, $[mRNA_i]$ é a concentração de referência do i-ésimo gene com $i \in \{1, 2, ..., n\}$, n representa o número total de genes analisados, $[mRNA_i]'$ é a concentração do i-ésimo gene no novo estado estável alcançado depois do estímulo ter sido aplicado.

Neste contexto, o coeficiente de co-controle expressa como duas variáveis mudam quando um único parâmetro é perturbado. E a força de regulação expressa o quanto de uma perturbação em uma variável é propagado para outra através de uma particular via. E redefine-se o coeficiente de co-controle,

$$^{V_m}O^i_j = \frac{\partial[mRNA_i]/[mRNA_i]}{\partial[mRNA_j]/[mRNA_j]},$$

na qual $^{V_m}O^i_j$ é o coeficiente co-controle definido para as concentrações de *mRNA* de dois genes i e j quando uma taxa de transcrição é perturbada.

A força de regulação do $mRNA_i$ no $mRNA_i$ é definida como,

$$^{V_j}R^j_i = \varepsilon^{V_j}_{mRNA_i} C^{mRNA_j}_{V_j} = \frac{\partial V_j/V_j}{\partial[mRNA_i]/[mRNA_i]} \frac{\partial[mRNA_j]/[mRNA_j]}{\partial V_j/V_j},$$

onde $\varepsilon^{V_j}_{mRNA_i}$ é a flexibilidade de V_j por $mRNA_i$, $C^{mRNA_j}_{V_j}$ é o coeficiente de concentração-controle da j-ésima reação de transcrição sobre $mRNA_j$.

Estes coeficientes medem o quanto uma perturbação na variável escolhida (concentração de *mRNA*) altera a concentração de outras variáveis dentro de uma determinada via de regulação.

Segue uma representação matricial da rede apresentada na Figura 2.3,

$$\begin{pmatrix} ^{V_a}R^A_A & ^{V_a}R^A_B & ^{V_a}R^A_C \\ ^{V_b}R^B_A & ^{V_b}R^B_B & ^{V_b}R^B_C \\ ^{V_c}R^C_A & ^{V_c}R^C_B & ^{V_c}R^C_C \end{pmatrix} = \begin{pmatrix} -0.9 & -0.7 & 0 \\ 0 & -0.6 & 0.3 \\ 0.2 & -0.5 & -0.6 \end{pmatrix}$$

A medida quantitativa obtida diretamente a partir de dados gerados em experimentos de *microarrays* é dada por

$$^{V_m}\widetilde{O}^i_j = \frac{\Delta mRNA_i/mRNA_i}{\Delta mRNA_j/mRNA_j} = \frac{(FR_i - 1)(FR_j + 1)}{(FR_j - 1)((FR_i + 1))}.$$

A matriz concentração-controle do i-ésimo gene é dada abaixo

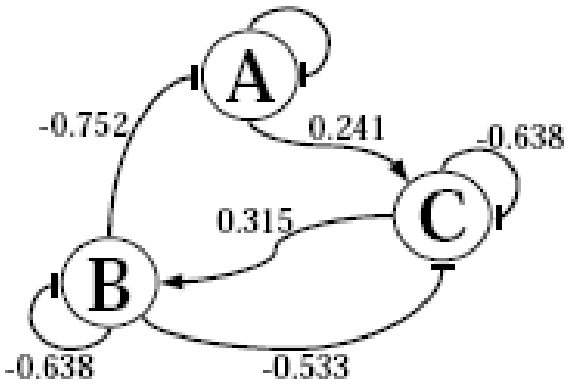

Figura 2.5: *Uma rede de regulação gênica formada por três genes. A interação representada por uma seta indica que o gene origem ativa o gene destino, enquanto que uma interação representada por uma barra indica que o gene origem reprime o gene destino. Os valores numéricos representam a força de regulação, valores negativos indicam inibição, 0 indica a ausência de influência direta e valores positivos indicam ativação.*

$$\widetilde{O}_i = \begin{pmatrix} {}^{V_1}\widetilde{O}_i^1 & \cdots & {}^{V_n}\widetilde{O}_i^1 \\ \vdots & \ddots & \vdots \\ {}^{V_1}\widetilde{O}_i^n & \cdots & {}^{V_n}\widetilde{O}_i^n \end{pmatrix}$$

Com esta metodologia é possível inferir as redes de regulação a partir de dados de expressão gênica. Este método requer a realização de perturbações em uma única taxa de transcrição e medir os níveis de mRNA de todos os genes depois de cada perturbação. Este processo é realizado para cada uma das taxas. Uma vantagem deste método é o fato dele ser quantitativo. Utilizando-o, não só procuramos as interações que estão presentes, mas também determinamos quão forte elas são, o que se reflete nas forças de regulação.

2.11 Características da modelagem de redes

Com o seqüenciamento do genoma humano tornou-se extremamente importante o desenvolvimento de ferramentas capazes de determinar as interações e funcionalidade dos genes. Neste sentido, a tecnologia de *microarray* pode ser empregada na análise de expressão gênica em grande escala. Redes de regulação gênica podem ser usadas para modelar interações gênicas, existindo atualmente na literatura uma ampla variedade de modelos. No entanto, na maioria dos casos não fica claro quais são os pontos fortes e fracos de cada abordagem, assim como suas similaridades e discordâncias. Algumas características sugeridas para comparar os modelos são: poder de inferência, poder de predição, robustez, consistência, estabilidade e custo computacional.

O processo de modelagem gênica começa com um modelo parametrizado que descreve o processo de regulação. Estes modelos são inspirados geralmente pela existência de conhecimento biológico sobre as interações entre dois genes, degradação de produto gênico, resposta de genes às entradas específicas, etc.

Além disso, conjuntos de dados de expressão gênica que contenham medidas de resposta a diferentes estímulos para genes de interesse ao longo do tempo podem ser usados para inferir os parâmetros desconhecidos do modelo.

Os modelos contínuos estudados na modelagem de redes de regulação gênica são classificados nas seguintes categorias: a) métodos baseados em comparações por pares (tais como: *Correlation Metric Construction* - CMC e redes de ativação e inibição); b) métodos que modelam redes preliminares (simples) de interações entre genes; c) métodos que modelam redes mais complexas que também descrevem produtos intermediários, tais como proteínas.

2.12 Modelo de redes *rough*

Estas redes modelam efeitos que resultam da combinação de diferentes genes de entrada, pela média da soma ponderada dos níveis de expressão. O termo *rough* refere-se ao fato da influência de todos os produtos intermediários serem resumidos em uma relação linear entre dois genes. Estes pesos fornecem informações sobre as relações entre genes, ou seja, peso 0 indica a ausência de interação e um peso positivo ou negativo corresponde a ativação ou repressão, respectivamente. O valor absoluto do peso indica a força da interação. Todo modelo de rede pode ser representado na seguinte equação diferencial generalizada

$$\frac{\partial x_i(t)}{\partial t} = R_i g\left(\sum_{j=1}^{J} W_{i,j} x_j(t) + \sum_{k=1}^{K} V_{i,k} u_k(t) + B_i\right) - \lambda_i x_i(t) \tag{2.2}$$

ou, de forma similar,

$$x_i[t+1] = R_i g\left(\sum_{j=1}^{J} W_{i,j} x_j[t] + \sum_{k=1}^{K} V_{i,k} u_k[t] + B_i\right) - \lambda_i x_i[t] \tag{2.3}$$

Nas Equações (2.2) e (2.3), $g(.)$ é uma função monótona que indica a regulação (ativação) da expressão, $x_i(t)$ e $x_i[t]$ representam a expressão gênica do i-ésimo gene no instante t, R_i a taxa constante do i-ésimo gene, $W_{i,j}$ a força do controle do j-ésimo gene no i-ésimo gene, $u_k(t)$ e $u_k[t]$ a k-ésima entrada externa no instante t, $V_{i,k}$ a influência da k-ésima entrada externa no i-ésimo gene, B_i o nível de expressão basal do i-ésimo gene, λ_i a constante de degradação do produto de expressão referente ao i-ésimo gene.

Algumas das metodologias para inferir os parâmetros dos modelos em geral são: *Expectation Maximization* (EM), *Simulated Annealing* (SA), *Genetic Algorithm* (GA), *Gradient Descent* (GD) e Regressão Linear (RL).

2.13 Redes complexas

Redes mais complexas não só modelam interações entre genes baseadas em medidas de níveis de mRNA, assim como modelam seus produtos intermediários e finais, tais como proteínas e metabólitos. Conseqüen-

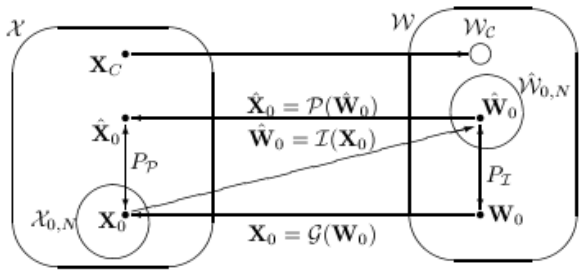

Figura 2.6: *Representação gráfica da estrutura que relaciona o espaço de expressão ($\mathcal{X}$) com o espaço da matriz de regulação gênica ($\mathcal{W}$).*

temente, seus parâmetros não podem ser inferidos a partir de conjuntos de dados que só contenham medidas de níveis de mRNA, pois é necessário conhecimento explícito sobre os níveis de concentração dos elementos intermediários.

As mudanças no nível de expressão de uma dada espécie química são modelados da seguinte forma:

$$\frac{\partial x_i}{\partial t} = \alpha_i \prod_{j \in S_I} x_j^{Wi,j} - \beta_i \prod_{k \in S_D} x_k^{i,k}$$

Diversas características fundamentais na modelagem de redes gênicas são: poder de inferência, poder de predição, robustez, consistência, estabilidade e custo computacional. Conseqüentemente, estas características são apropriadas para avaliar e comparar as aproximações feitas pelas diferentes abordagens. As definições das mesmas, assim como as suas motivações serão apresentadas seguindo a Figura 2.6, que apresenta a estrutura e notação empregada.

Para uma particular classe de modelos de redes gênicas, $\mathcal{W}$ representa o espaço de matrizes de regulação gênica, sendo W uma realização de uma matriz de regulação. $\mathcal{X}$ denota o espaço do conjunto de dados associado com $\mathcal{W}$, sendo X um particular conjunto de dados. Por propósito experimental uma particular matriz, W_0, é escolhida para representar uma rede "real". Baseado em W_0 e condições iniciais, um conjunto de dados (X_0) é gerado ($X_0 = (W_0)$). É determinada uma estimativa de W_0 que só depende de X_0, e esta é representada por $\widehat{W}_0$ e pode ser determinada da seguinte forma: $\widehat{W}_0 = \mathcal{I}(X_0)$. $\widehat{W}_0$ pode ser empregado para fazer uma predição do dado de expressão, dadas as mesmas condições iniciais. Isto produz uma aproximação de X_0, denotada por $\widehat{X}_0$, sendo $\widehat{X}_0 = (\widehat{W}_0)$.

O poder de inferência, $P_{\mathcal{I}}$, é medido como a similaridade entre as matrizes de regulação gênica real (W_0) e a estimada ($\widehat{W}_0$) e é definido por

$$P_{\mathcal{I}}(W_0, \widehat{W}_0) = 0.5(1 + \rho(W_0, \widehat{W}_0)),$$

na qual $\rho(.,.)$ é o coeficiente de correlação de Pearson.

O poder de predição, $P_{\mathcal{P}}$, é refletido na exatidão da predição, ou seja, quanto $\widehat{X}_0$ aproxima-se de X_0. Para

modelos de redes, matrizes de regulação são estimadas por minimização do erro na predição. Conseqüentemente, isto é importante para ganhar informação na relação entre $P_{\mathcal{P}}$ e $P_{\mathcal{I}}$ para os diferentes modelos na comparação. Esta medida é definida por

$$P_{\mathcal{P}} = \frac{1}{(1 + E_{MSE})},$$

na qual o erro quadrático médio (E_{MSE}) é dado por

$$E_{MSE}(\mathrm{X}_0, \widehat{\mathrm{X}}_0) = \frac{1}{TN} \sum_{i,t} (X_0(i,t) - \widehat{X}_0(i,t))^2.$$

2.14 Robustez

Ruído é um fenômeno que sempre está presente e que necessariamente precisamos lidar com ele. Em particular, medidas de expressão gênica são ruidosas e, conseqüentemente, isto é importante para conhecer em que grau uma matriz de regulação gênica exata será extraída na presença de ruído.

Seja $\mathcal{X}_{0,N} \subset \mathcal{X}$ um grupo de conjuntos de dados de expressão que são obtidos quando o ruído, com distribuição $N(0, \sigma^2)$, é incorporado a X_0, ou seja, $\mathcal{X}_{0,N} = \{\mathrm{X}_{0,\sigma} | \mathrm{X}_{0,\sigma} = \mathrm{X}_0 + N(0, \sigma^2), \sigma \in [0, \sigma_{MAX}]\}$. Seja $\widehat{\mathcal{W}}_{0,N}$ o conjunto de matrizes de regulação que são obtidas através de inferência em cada elemento de $\mathcal{X}_{0,N}$, ou seja, $\widehat{\mathcal{W}}_{0,N} = \{\widehat{\mathrm{X}}_{0,\sigma} | \widehat{\mathrm{W}}_{0,\sigma} = \mathcal{I}(\mathrm{X}_{0,\sigma}), \mathrm{X}_{0,\sigma} \in \mathcal{X}_{0,N}\}$. Uma medida de robustez, $P_{\mathcal{R}}$ é definida como a correlação mínima entre as matrizes de regulação estimadas em $\widehat{\mathcal{W}}_{0,N}$,

$$P_{\mathcal{R}} = \min_{\left\{\widehat{\mathrm{W}}'_{0,\sigma}, \widehat{\mathrm{W}}_{0,\sigma}\right\} \in \widehat{\mathcal{W}}_{0,N}} \left(0.5(1 + \rho(\widehat{\mathrm{W}}_{0,\sigma}, \widehat{\mathrm{W}}'_{0,\sigma}))\right).$$

2.15 Consistência

Uma das propriedades que atrapalham a maioria dos dados de expressão gênica disponíveis é o número relativamente grande de genes comparados com o número de instantes de tempos nos quais foram feitas as medições. Isto é conhecido como "problema de dimensionalidade" e é uma causa importante de inconsistência no modelo de rede gênica inferido, sendo desta forma uma importante característica para investigar.

Um modelo de redes gênicas é dito ser inconsistente se múltiplos conjuntos de parâmetros podem ser inferidos a partir dos mesmos dados de expressão. Para um conjunto de dados de expressão arbitrário, $\mathrm{X}_C \in \mathcal{X}$, definimos o seguinte conjunto de matrizes de regulação gênica,

$$\mathcal{W}_C = \{\mathrm{W}_C | \mathrm{W}_C = \mathcal{I}(\mathrm{X}_C), P_{\mathcal{I}}(P(\mathrm{W}_C), \mathrm{X}_C) > 1 - \epsilon\}.$$

O grau de consistência de um modelo é dado por:

$$P_C = \min_{\{W_{C,i}, W_{C,j}\} \in \mathcal{W}_C} \left(0.5(1 + \rho(W_{C,i}, W_{C,j}))\right).$$

2.16 Estabilidade

Devido a energia limitada capacidade e armazenamento na célula, as concentrações de produtos de expressão gênica, tais como mRNA permanecem limitadas. Todas as redes gênicas reais são conseqüentemente estáveis por definição. Desta forma, modelos de redes gênica inferidos também deve ser estáveis em relação ao real.

Um modelo parametrizado por uma específica matriz de regulação gênica é estável se os níveis de expressão estimados permanecem sempre limitados por algum estado inicial finito. Se conjuntos de treinamento infinitamente grandes estão disponíveis, e E_{MSE} permanece limitado, então a estabilidade pode ser garantida. Em conjuntos de dados finitos, a exigência para que E_{MSE} seja limitado precisa ser aumentada por outros requisitos.

2.17 Custo computacional

Métodos que precisam de tempos de processamento extremamente longos até determinar uma solução são obviamente indesejáveis, a menos que o tempo de espera adicional resulte em resultados significativamente melhores. Métodos que não podem calcular soluções analíticas e empregam soluções iterativas aproximadas requerem, em geral, tempos de processamento maiores. Avaliações empíricas podem ser empregadas para obter uma estimativa do tempo requerido para obter uma solução para cada uma das aproximações.

2.18 Alguns tipos de redes e suas propriedades

Modelagem matemática pode prover a estrutura conceitual necessária para integrar os fenômenos estudados neste contexto em um ambiente coerente da operação de regulação gênica; identificar princípios de desenho da arquitetura bioquímica de sistemas gênicos; entender as respostas de organismos normais e mutantes aos estímulos; verificar a consistência e a completude dos conjuntos de reações assumidas para modelar o sistema específico.

Algumas aproximações comumente usadas para modelar sistemas de regulação gênica são: o controle é realizado ao nível de transcrição; a produção protéica é um processo contínuo, com a taxa determinada por um balanço de ativação gênica *vs* repressão.

Algumas considerações para a aproximação booleana são: o estado de cada gene pode ser representado como ligado (1) ou desligado (0); o controle de regulação de expressão gênica pode ser aproximado por regras de lógica booleana; comumente é assumido que todos os elementos atualizam seus estados (1,0) ao mesmo tempo, ou seja, são considerados tempos entre instantes o suficientemente grandes para os elementos mudarem de estado dependendo dos estados no instante anterior.

As atualizações podem ocorrer de duas formas, sincronizadas ou não. No caso das sincronizadas, todos os elementos trocam de estado ao mesmos tempo. Enquanto que nas atualizações não sincronizadas, os elementos não trocam de estado ao mesmo tempo. Estas últimas não são comumente consideradas.

Ciclos de *feedback* negativos são importantes para manter *homeostasis* em níveis de produtos gênicos. Enquanto que ciclos de *feedback* positivos são importantes para permitir múltiplos estados estáveis de níveis de produto gênico, ou seja, múltiplas redes com estados estáveis.

Abordagens para modelar sistemas gênicos com lógica booleana, equações diferenciais ou modelos híbridos discreto-contínuo podem ser usadas.

O modelo *operon*, para o caso em particular de um gene estrutural, pode ser representado pelo seguinte sistema de equações diferenciais,

$$\begin{aligned} \tfrac{\partial x_1}{\partial t} &= f(x_n) - k_1 x_1 \\ \tfrac{\partial x_j}{\partial t} &= x_{j-1} - k_j x_j, \qquad j = 2, 3, ..., n, \end{aligned} \tag{2.4}$$

nas quais x_i é a concentração do produto do i-ésimo gene.

Segue outro tipo de modelo freqüentemente usado para representar redes gênica,

$$\frac{\partial x_i}{\partial t} = a_i(Z) - b_i(\lambda)x_i, \qquad j = 1, 2, ..., n,$$

na qual Z é o vetor vetor de funções limite *sloped stepped* ou *steeply*, a_i freqüentemente apresenta uma série de passos em hiper-superfícies diferentes dentro do espaço de x_i, λ vetor de constantes, embora possa expressar a cinética hiperbólica ou *sigmoidal* da degradação do produto gênico, Z_i com i=1,2,...,n são as componentes do vetor Z, $Z_i = S_i(x_1, x_2, ..., x_n, \eta_i)$, S_i denota uma função limitada por η_i.

Note que se os b_i são constantes e os a_i são tomados como funções de x_{i-1}, com exceção de $i = 1$, então obtém-se um caso específico da Equação (2.4).

Comparação entre as aproximações booleana e contínua: (i) a aproximação contínua é mais cara computacionalmente; (ii) freqüentemente aproximação contínua é preferida por ser mais exata; (iii) embora um estado estável da representação booleana de um sistema gênico possa corresponder qualitativamente a um estado estável análogo em uma representação contínua com equações diferenciais ordinárias, o recíproco não é verdadeiro; (iv) nem todo estado estável da representação contínua é solução estável na representação booleana; (v) soluções periódicas na representação booleana podem não corresponder com soluções periódicas no sistema contínuo.

Podemos imaginar um exemplo onde uma rede formada por dois genes (X_1, X_2), na qual X_1 ativa X_2 e X_2 reprime X_1. Empregando representação booleana temos quatro estados possíveis: $(0, 0), (0, 1), (1, 0)$ e $(1, 1)$. Considerando atualização sincronizada, verifica-se que temos uma solução periódica formada pelos quatro estados possíveis: $(1, 1) \to (0, 1) \to (0, 0) \to (1, 0) \to (1, 1)$. A representação contínua deste sistema usando funções *sigmoidal steep* para ativação e repressão não tem oscilação estável.

2.19 Modelos de redes neurais

Redes neurais pode ser empregada na modelagem de redes de regulação gênica. Neste caso a entrada de regulação para o i-ésimo gene é dada por

$$g_i = \sum_j w_{ij} y_j + \theta_i,$$

onde w_{ij} é a força da interação de regulação entre os genes i e j.

O nível de ativação para um determinado gene é calculado pelas entradas de regulação e degradação do mesmo,

$$\frac{\partial x_i}{\partial t} = \alpha_i f_i(g_i) - \gamma_i x_i,$$

na qual α e γ são as taxas de ativação e degradação e f_i a função de transferência *sigmoid*.

2.20 Grafos direcionados ou não

Uma forma natural de modelar redes de regulação gênica é olhá-la como um grafo direcionado [Jong, 2002]. Um grafo direcionado G é definido pelo par (V, E), no qual V é um conjunto de vértices e E um conjunto de elos. Um elo direcionado é formado por um par de vértices (i, j), no qual i representa a origem do elo e j o fim. Os vértices do grafo direcionado correspondem a genes ou outros elementos de interesse no sistema de regulação, enquanto os elos denotam as interações entre os elementos. A representação gráfica de uma rede de regulação pode ser generalizada de diversas formas. Tantos os vértices quanto os elos podem ter uma descrição, como por exemplo, informação dos genes da rede, assim como informação das interações entre eles. Por definição, um elo direcionado é denotado pelo trio (i, j, s), no qual s pode ser "+" ou "-", indicando ativação ou repressão respectivamente. Então, os elos podem ser representados pelo trio (i, J, S), no qual i é o elemento origem, J uma lista de elementos reguladores e s uma lista indicando o tipo de influência em cada caso.

3
Simulated Annealing - SA

ESTA METODOLOGIA [van Laarhoven & Aarts, 1987] tem seu nome associado ao processo de recozimento (arrefecimento simulado), termo usado na metalurgia. A idéia principal do processo físico é: 1) a temperatura do sólido é aumentada para um valor máximo, no qual ele se funde; 2) resfriamento lentamente até que o material se solidifique.

É definida a função objetivo E que queremos minimizar, o que é similar a energia interna do material no processo físico. $E(S)$ corresponde à energia do estado S. Também é definida a temperatura T, a qual inicia em um valor alto e depois é diminuido lentamente. O objetivo do SA é diminuir a energia na medida que diminui a temperatura.

3.1 Heurística do SA:

Entradas: T_0, T_f e α;
$T = T_0$;
$S_i = S_0$, estado inicial qualquer;
enquanto $T > T_f$:
    S_j = seleção de um estado vizinho de S_i;
    Se $(\triangle E_{ij} \leq 0)$ ou $\left(U(0,1) \leq p(\triangle E_{ij}) = \exp^{-\triangle E_{ij}/T}\right)$ Então $S_i = S_j$;
    $T = \alpha T$;
fim-enquanto.

T_0 é a temperatura inicial, esta deve ser o suficientemente alta para que soluções ruins sejam aceitas no inicio com bastante frequencia. T_f é a temperatura final e deve ser bem próxima de 0, a parte final do algoritmo é similar a um método de descida, pois soluções piores têm probabilidade quase 0 de serem aceitas, ou seja, apenas aceitam-se soluções que diminuim a energia. $\exp^{-\triangle E_{ij}/T}$ é chamado na literatura de fator de Boltzmann. S_i refere-se ao estado i e $E(S_i)$ denota a energia de S_i. $\alpha \in (0,1)$ é a proporção de redução de temperatura. Os possíveis casos no laço são: a) $\triangle E_{ij} < 0$, a nova solução é melhor (redução de energia) e o método aceita a nova solução; b) $\triangle E_{ij} = 0$, caso de estabilidade, esta situação é pouco provável e sua aceitação ou não é indiferente; c) $\triangle E_{ij} > 0$, solução pior (aumento de energia) com a ceitação mais

provável a altas temperatura.

Seja $\{X_n\}_{n\geq 0}$ uma cadeia de Markov, com matriz de transição $P_{ij} = (X_k = j|X_{k-1} = i)$. Também podemos definir, $a_i(k)$ sendo a probabilidade de estar no estado i no instante k,

$$a_i(k) = (X_k = i) = \sum_l a_l(k-1)_{li}(k-1,k), \text{ k=1,2,...}$$

A matriz de transição P_{ij} também pode ser descomposta da seguinte forma:

$$P_{ij} = \begin{cases} G_{ij}A_{ij}, & \forall j \neq i \\ 1 - \sum_{l\neq i} G_{il}A_{il}, & j = i \end{cases}$$

onde G e A são denotadas como as matrizes de geração e aceitação, respectivamente. P é a matriz estocástica, $\sum_j P_{ij} = 1, \forall i$.

A seguir são apresentados dois teoremas, os quais garantem a convergência assintótica do algoritmo para o mínimo global sob certas condições para as matrizes G e A.

3.1.1 Teorema 1

A distribuição estacionária (q) de uma cadeia de Markov homogenea finita existe se a cadeia de Markov é irredutível e aperiódica.

Temos que $\forall i$, $q_i = \sum_j q_j P_{ji} > 0$ e $\sum_i q_i = 1$. A cadeia é irredutível se e somente se $\forall(i,j)$ $\exists n$: $1 \leq n < \infty$, tal que $(P^n)_{ij} > 0$. A cadeia também é aperiódica se e somente se o máximo divisor comum $\forall n \geq 1$ tal que $(P^n)_{ii} > 0$ é 1.

Podemos definir $(S_i) = q_i(c) = \frac{1}{Q(c)} \exp(-\frac{E(S_i)}{c})$, onde $Q(c)$ é uma constante de normalização e c é um parâmetro de control (exemplo: a temperatura).

Seja $\psi(\gamma, c)$ uma função positiva com dois argumentos tal que: 1) $\forall S_i \in R, c > 0$: $\psi(E(S_i), c) > 0$, onde R representa o conjuntos de estados possíveis (espaço de estados); 2) $\forall S_j \in R$, $\sum_{i\neq j} \psi(E(S_i), c)G_{ij}(c)A_{ij}(c) = \psi(E(S_j), c) \sum_{i\neq j} G_{ji}(c)A_{ji}(c)$.

$\lim_{c\downarrow 0} q(c) = \pi$ (medida estacionária) se: 1) $\lim_{c\downarrow 0} \psi(\gamma, c) = \{0, \gamma > 0; \infty, \gamma < 0\}$; 2) $\frac{\psi(\gamma_1,c)}{\psi(\gamma_2,c)} = \psi(\gamma_1 - \gamma_2, c)$; 3) $\forall c > 0 : \psi(0, c) = 1$.

3.1.2 Teorema 2

Se os dois argumentos da função $\psi(E(S_i) - E_{opt}, c)$ são escolhidos como $A_{i_0 i}(c)$ para uma configuração arbitrária $S_{i_0} \in R_{opt}$ e se $G(c)$ e c são independentes então a distribuição estacionária,

$$q_i(c) = \frac{A_{i_0 i}(c)}{\sum_{j\in R} A_{i_0 i}(c)}, \forall S_i \in R,$$

tal que $A(c)$ e G satisfazem: 1) $\forall(S_i, S_j) \in R$: $G_{ji} = G_{ij}$; 2) $\forall(S_i, S_j, S_k) \in R$: $E(S_i) \leq E(S_j) \leq E(S_k) \Rightarrow A_{ik}(c) = A_{ij}(c)A_{jk}(c)$; 3) $\forall(S_i, S_j) \in R$: $E(S_i) \geq E(S_j) \Rightarrow A_{ij}(c) = 1$; 4) $\forall(S_i, S_j) \in R, c > 0$: $E(S_i) < E(S_j) \Rightarrow 0 < A_{ij}(c) < 1$. R_{opt} refere-se ao conjunto de mínimos globais e $E_{opt} = E(S_{i_0}) = \min_{S_i\in R} E(S_i)$.

4

Modelagem de vias de sinalização

DE MODO GERAL, as trajetórias no tempo dos níveis de expressão, para todos os genes de um organismo ou parte deles, podem ser modeladas como funções aleatórias no tempo. Uma das abordagens mais empregadas tem sido redes booleanas, as quais podem ser muito bem adaptadas a estudos do genoma. Como realizar a modelagem usando redes booleanas a partir de dados de *microarray* ainda é um problema em aberto. Uma das vantagens de usar essas redes é o fato de ser possível incorporar o conhecimento biológico que se tem a priori. Desta forma, determinam-se os relacionamentos regulatórios, os quais serão refletidos pelas classes de funções booleanas dos genes, representando o conhecimento a priori. Umas das possíveis abordagens para tirar as informações é a de agrupamento, podendo esta ser supervisionada ou não, considerando as dependências e independências existentes em dados de *microarray*. A técnica de agrupamento tem sido usada para obter a similaridade existente nos perfis de expressão.

Uma via de sinalização pode ser descrita de forma simplificada como os fluxos de informação e os componentes (proteínas, genes e pequenas moléculas) que formam uma via de sinalização, podem ser simplificados em uma representação de grafo orientado, onde os nós são os componentes e os elos são as interações. Cada elo do grafo representa uma interação entre dois componentes, o componente destino é o componente dependente e os componentes origens são os que influenciam no componente destino. Os elos carregam informação referente ao tipo de interação: +1 e -1, que representam ativação ou repressão, respectivamente.

Definição de uma rede de regulação gênica: Seja $G(V, F)$ uma rede de regulação gênica determinística, na qual $V = \{x_1, \ldots, x_n\}$ é o conjunto de nós representando os genes e $F(f_1, \ldots, f_n)$ é uma lista de funções. De forma geral, cada nó toma valores de um conjunto finito, sendo os mais comumente usados $\{0, 1\}$ e $\{-1, 0, 1\}$. Uma questão que requer singular atenção não é só a reconstrução da rede como um todo, mas também poder identificar as sub-redes ativas, ou seja, regiões conectadas da rede que apresentam mudanças significativas da expressão, dado determinado conjunto de condições.

4.1 Abordagem discreta (rede booleana)

Dada uma determinada via, podemos definir $g = \{1, 2, ..., m, ..., n\}$, sendo m o número de genes da rede medidos no experimento e n o número total de genes na rede. Também defini-se o grafo, $G = (g, \varepsilon)$, no qual ε representa as arestas orientadas, sendo o peso das mesmas (+ ou -), que representa (ativação ou repressão). Neste momento, estamos estudando a abordagem booleana, e podemos considerar a rede básica como $B = \{G, \psi\}$, onde ψ é a função controle.

A função controle é definida pelos critérios biológicos, considerando todos os genes envolvidos na rede e as interações que existem entre eles. Para cada componente, i, da rede podemos associar um conjunto de genes que o regulam, I_i. Da mesma forma, podemos também definir o conjunto de genes que são regulados por ele, O_i. A partir da função controle podemos obter, a a partir da configuração dos genes em I_i, qual é a tendência qualitativa instantânea do i-ésimo gene. Onde esta tendência, $\psi = \psi(1), ..., \psi(n)$ pode ser (-1,0,+1), repressão, ficar constante e ativação, respectivamente.

Se pensarmos em uma série temporal com k instantes de tempos, podemos definir $x_i(t)$, como o nível de expressão para o i-ésimo genes no instante t. Associado a cada gene, também temos θ_i, isto assumindo o caso booleano. Estes parâmetros são empregados para determinar o estado qualitativo do gene, dado o respectivo x_i. Note que só precisamos de um θ para cada gene, valor que define a partir dele ativo e abaixo não.

Neste caso, podemos definir o espaço de configurações possíveis para a rede,

$$\Sigma = \mathrm{X}_{i \in g}\{0, 1\},$$

Sendo, neste caso, a função controle denotada como,

$$\psi : \Sigma \rightarrow \{-1, 0, +1\}.$$

Em particular, foi estudado um módulo (subgrupo de elementos e interações, Figura 4.1) da rede Wnt (http://stke.sciencemag.org/cgi/cm/stkecm;CMP_5533). Nesta sub-rede foi definido o grafo que a forma, onde os nós e os elos representam os genes e suas interações, respectivamente.

Inicialmente foi feito um experimento começando de possíveis configurações escolhidas aleatoriamente e simulando um grande número de iterações. Isso nos levou a 22 configurações, que eram os que apareciam com mais freqüência. Assumimos atualização sincronizada (todas as configurações são atualizadas ao mesmo tempo) e as regras de atualização são comuns para todos os genes. Uma continuação imediata deste trabalho é a mudança destas regras por outras mais específicas e com mais argumentos biológicos. Depois, foi realizado um experimento com 22 simulações, onde cada uma delas começa em uma das configurações determinadas anteriormente. É interessante destacar que as 22 configurações geram um laço fechado (Figuras 4.2 e 4.3) onde, uma vez nele, a sub-rede não consegue mais sair. Um terceiro experimento foi feito

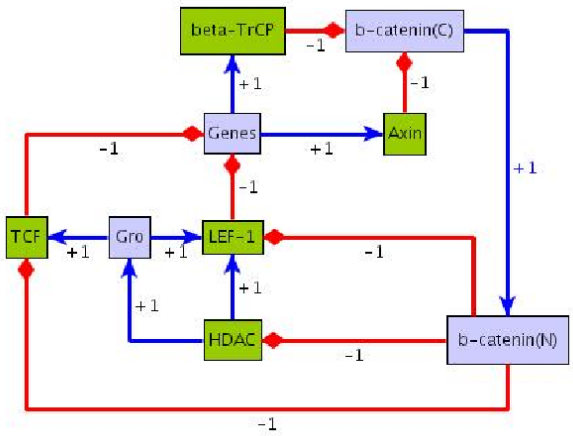

Figura 4.1: *Subrede da Wnt escolhida para gerar as simulações, onde "+1" (seta) indica ativação e "-1" (diamante) indica repressão.*

Índice do gene	Descrição
1	beta-TrCp
2	Genes
3	LEF-1
4	HDAC
5	Gro
6	TCF
7	b-catenin(C)
8	Axin
9	b-catenin(N)

Tabela 4.1: *Lista dos genes que formam o módulo da Wnt em estudo, apresentado na Figura* 4.1.

que consistia da realização de uma simulação iniciada a partir de cada uma das configurações possíveis (3^9 = 19683), e o resultado foi muito interessante. O número máximo de iterações que demora uma simulação, iniciando-a em qualquer estado possível, para atingir o laço dos 22 é 17 iterações; isto excluindo-se 9 configurações estacionários, mostradas na Figura 4.4. A numeração dos genes das Figuras 4.2 e 4.4 está na Tabela 4.1.

4.2 Abordagem contínua (equações diferenciais)

A modelagem com equações diferenciais começa com o critério a priori, que já se tem da via (rede de regulação). Nela, são representados os genes presentes, bem como suas interações para realizar as funções necessárias. As interações podem ser muito simples, por exemplo, quando um gene sozinho, ativa outro gene. Mas também pode ser muito complexa, por exemplo quando um grupo de genes (3, 4 ou mais) ativam ou reprimem um outro gene, levando em consideração que, a junção entre os genes que regulam pode ser muito variada. Desta forma, definimos as derivadas dos genes na via no sistema de equações diferenciais,

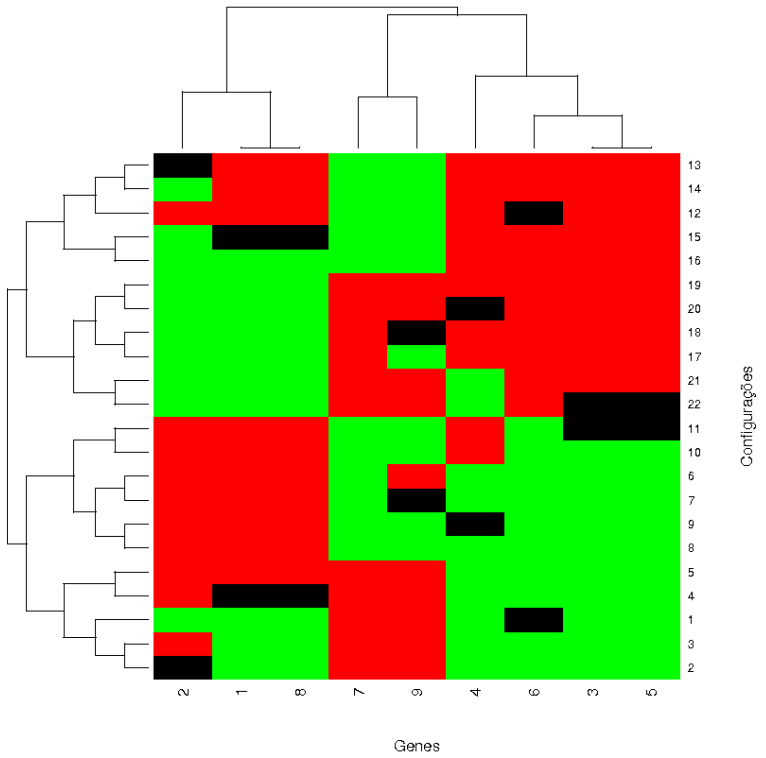

Figura 4.2: *Heatmap para as 22 configurações que geram o laço estacionário. A identificação dos genes é apresentada na Tabela* 4.1.

$$\dot{x}_i^t = f_i(x^t) - g_i(x^t)$$

onde f_i e g_i indicam as velocidades de produção e degradação, respectivamente, do gene i, o vetor x^t representa os níveis de expressão gênica para os genes na via no instante t, $f_i(x^t) = (1 - x_i^t)\lambda_i(x^t)$, $g_i(x^t) = x_i^t \mu_i(x^t)$, $\lambda_i(x^t) = \exp(\sum \alpha_{ji} x_j^t + a_i)$, $\mu_i(x^t) = \exp(\sum \beta_{ji} x_j^t + b_i)$. α_{ji} e β_{ji} representam as forças da interação do gene j com o gene i, a_i e b_i podem ser interpretadas como um tipo de dependência externa (estado basal) para o gene i. Segue um exemplo: gene A ativa o gene B,

$$\dot{x}_B^t = \exp(\alpha_{AB} x_A^t + a_B) - \left(\exp(\alpha_{AB} x_A^t + a_B) + \exp(\beta_{AB} x_A^t + b_B)\right) x_B^t$$

A via EGF do Stke (http://stke.sciencemag.org/cgi/cm/stkecm;CMP_14987), além de estar associada ao câncer, também fica relativamente bem coberta (genes que a compõem estão medidos no experimento do qual temos os dados reais). Uma vez escolhido o módulo da EGF, que será modelado, é escolhido um sistema de equações diferencias (como o exemplo anterior) para ajustar este modelo. Empregando o nosso algoritmo de estimação, implementado nos aplicativos R e Xppaut, com a ideía de SA (Seção 3), o modelo é ajustado. Este ajuste é realizado gerando várias iterações até convergir para um conjunto de parâmetros, os quais ajustam o modelo, obtendo o menor erro quadrático (ou alguma outra medida de erro entre as séries estimadas e reais). Uma vez realizada a simulação, os parâmetros obtidos são os nossos estimadores. Esta idéia de modelagem, proposta pelo Prof. Dr. Eduardo J. Neves [Fernandez *et al.*] faz muito sentido e

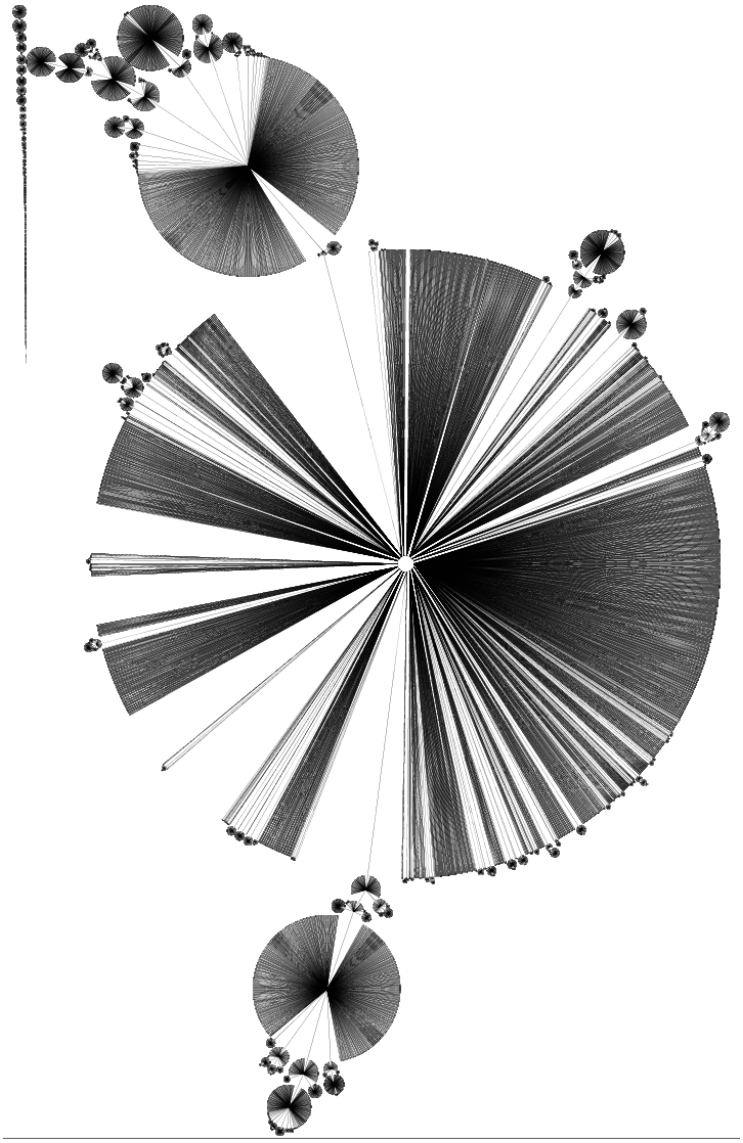

Figura 4.3: *Grafo que representa a dinâmica do módulo da Wnt.*

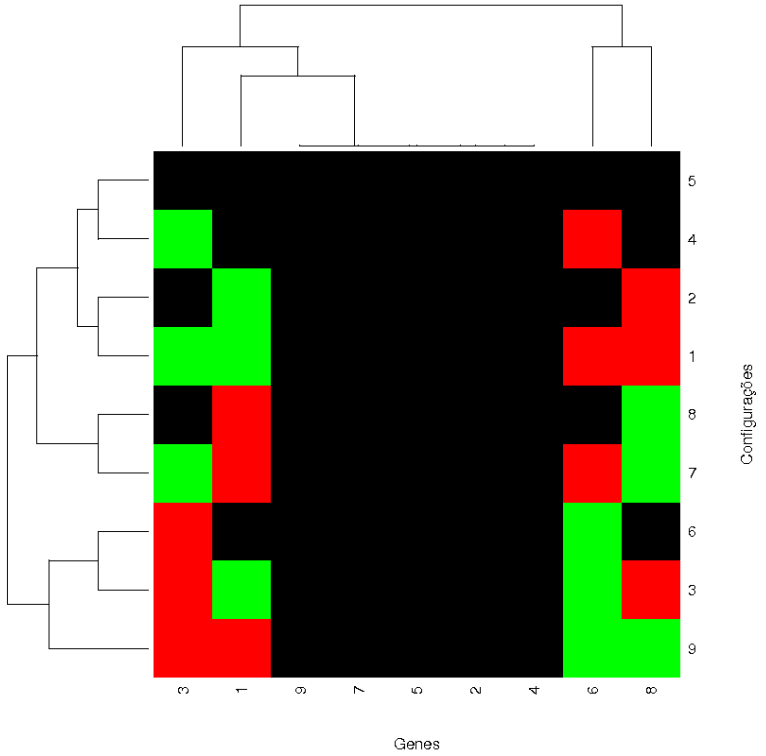

Figura 4.4: *Heatmap para as 9 configurações estáticas. A identificação destes genes é apresentada na Tabela* 4.1.

apresenta bons resultados em dados simulados, também foi testada em dados reais, ambos os resultados são apresentados neste trabalho. Um desafio é o fato de termos da via apenas alguns genes dela na lâmina usada no experimento.

5
Contribuições deste trabalho

A PRINCIPAL CONTRIBUIÇÃO DESTE TRABALHO consiste em abordar a modelagem de redes de regulação gênica, não apenas com dados simulados, como também com dados reais. Neste trabalho, propomos uma metodologia, algoritmo e ferramenta desenvolvida em R (http://www.r-project.org) e Xppaut (http://www.math.pitt.edu/ bard/xpp/xpp.html), para modelagem de redes de regulação gênica, a partir de dados de expressão gênica (experimentos com *microarrays*) empregando sistema de equações diferenciais. Neste caso, somente são especificadas as derivadas (sistema de equações diferenciais, taxa de ativação e repressão) para cada um dos genes na via que não seja entrada (forçado). Uma das maiores dificuldade em lidar com este problema de modelagem em dados reais, é a quantidade de dados ausentes, ou seja, muitas vezes em experimentos biológicos com *microarray* não são estudados (medidos) todos os genes para determinadas vias de interesse. Esta falta de informação e o grande número de parâmetros envolvidos tornam o problema mais difícil. Esta deve ser uma das causas pelas quais o número de trabalhos na literatura [Brown & Sethna, 2003; Haberichter *et al.*, 2007; Kim *et al.*, 2007; Legewie *et al.*, 2006], em modelagem de vias empregando dados reais seja mínimo.

5.1 Algoritmo de estimação empregando SA para dados reais e simulados

① Todos os parâmetros (α's, β's, a's, b's, e x_0's não medidos) são inicializados aleatoriamente;

② É gerado o sistema de equações diferenciais do modelo em formato Xppaut (.ode) com as derivadas, condições inicias e parâmetros;

③ O modelo é armazenado na memoria RAM do computador;

④ O aplicativo R roda o Xppaut por uma chamada do sistema operacional e o resultado obtido para o modelo é armazenado também na RAM;

⑤ O resultado para os genes medidos nos instantes observados é comparado com os dados reais, obtendo-se $H_{novo} = \sum_{i=1}^{n_{medidos}} (\max((serie_{estimada} - serie_{real})^2))$;

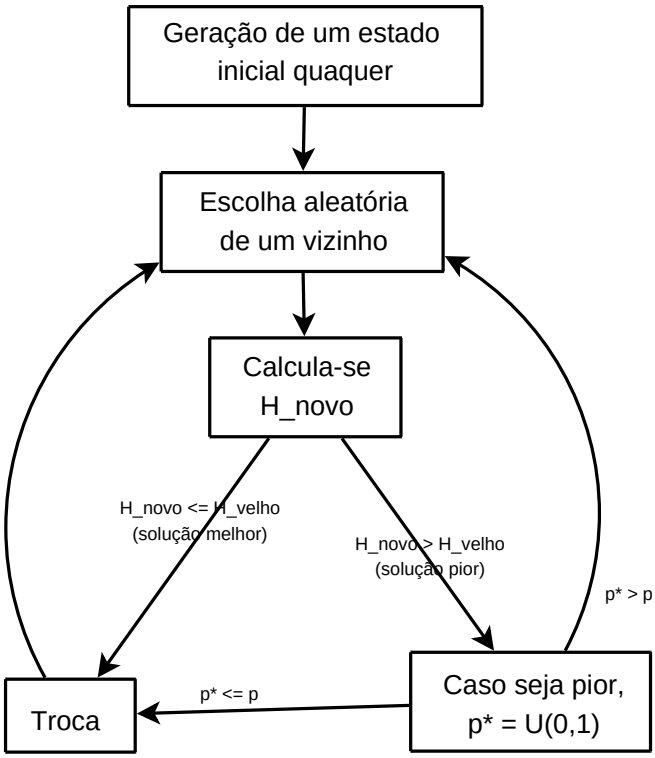

Figura 5.1: *Diagrama do algoritmo SA, Seção* 3.

⑥ São escolhidos alguns parâmetros aleatoriamente (aproximadamente 20% deles) e gera-se pequenas mudanças, $N(0, \sigma^2)$, para eles;

⑦ Com os novos parâmetros, repete-se 2, 3, 4 e 5;

⑧ Se $H_{novo} < H_{velho}$, aceita-se a troca e são atualizados os parâmetros para os quais geramos mudanças. Caso, $H_{novo} >= H_{velho}$, então só aceita-se a troca com probabilidade p;

⑨ Repete-se 6, 7 e 8, N vezes.

Segue um diagrama resumido do algoritmo SA,

5.1.1 Observações:

① N é o suficientemente grande para o sistema percorrer várias propostas e esfriar lentamente, grantindo assim que a estimativa calculada seja mínimo local;

② Foi escolhido para a transferência dos arquivos a RAM e não o HD, por ser a primeira aproximadamente três vezes mais rápida;

③ σ é o suficientemente grande para que o sistema possa explorar áreas maiores;

④ p no começo é o suficientemente grande para que o sistema possa explorar outras soluções (embora piores), e este vai diminuindo até no final ser o suficientemente pequeno, tal que só aceita trocas se $H_{novo} < H_{velho}$, até esfriar totalmente;

⑤ p depende do índice da iteração e da proporção H_{velho}/H_{novo}, enquanto menor seja esta proporção menor será a chance de troca.

5.2 Algoritmo de estimação empregando SA: versão modular

No caso em que a via seja muito grande (por exemplo, mas do que 20 genes), este algoritmo pode não ser viável pelo fato de que o número de parâmetros envolvidos é muito grande e o tempo computacional para rodar o *simulated annealing* seria muito grande. Então desenvolvemos uma versão modular deste algoritmo. A idéia básica é a seguinte: primeiro identifica-se os módulos que formam a via; depois roda-se o algoritmo anterior para um dos módulos e em um processo iterativo vão se incorporando os outros módulos até termos a via completa. Este algoritmo de estimação modular empregando SA, para dados reais e simulados tem a seguinte característica. Quando começamos por um módulo menor do que a via completa, fazemos uma busca mais detalhada das possíveis soluções e entre elas podemos escolher a melhor (menor H). Em cada iteração, quando incorporamos um módulo novo, tentamos manter as estimativas já calculadas: i) inicializamos a busca com elas; ii) é atribuída uma probabilidade menor de escolha para troca nelas; iii) caso sejam escolhidas, propomos mudanças menores para elas. Para os parâmetros associados ao módulo que está sendo incorporado se faz uma busca similar à que foi realizada no módulo inicial.

Podemos separar os genes em entradas (aqueles que nenhum dos outros influenciam nele) e não entrada (pelo menos um gene influência nele). Também podemos dividir os genes em outros dois grupos, medidos (aqueles para os quais temos os níveis de expresão reais) e não medidos (aqueles que não foram incluidos no experimento). Se cruzamos estas duas divisões podemos gerar quatro grupos: os entrada medidos, os entrada não medidos, os não entrada medidos e os não entrada não medidos. Desta forma, não serão considerados os genes que sejam entrada e que não sejam medidos pelo fato de não aportar informação útil, os outros três tipos sim serão considerados. Dos entrada medidos não conseguiremos gerar as equações diferenciais, mas sim temos a evolução dos níveis de expressão reais ao longo do tempo, porém não poderemos usar esta informação para avaliar a qualidade do ajuste pois pelo fato de não termos as equações diferenciais para eles não será possível gerar os níveis de expressão simulado ao longo do tempo. Por outro lado os não entrada sim terão equação diferencial e consequentemente níveis de expressão simulado, mas para a avaliação da qualidade do ajuste apenas poderão ser considerados os não entrada medidos pois para os não entrada não medidos não temos os níveis de expressão reais.

Seguem alguns resultados com dados simulados. Foi eslcolhido para o experimento, um sub-módulo da via EGF, Figura 5.2. O nó em branco representa um gene não medido (não foi considerado no experimento biológico). Neste exemplo consideraremos o gene EGFR (medido) como forçado (entrada do sistema), ou seja, os pontos observados para este gene são entradas fixas ao sistema e os pontos não observados são

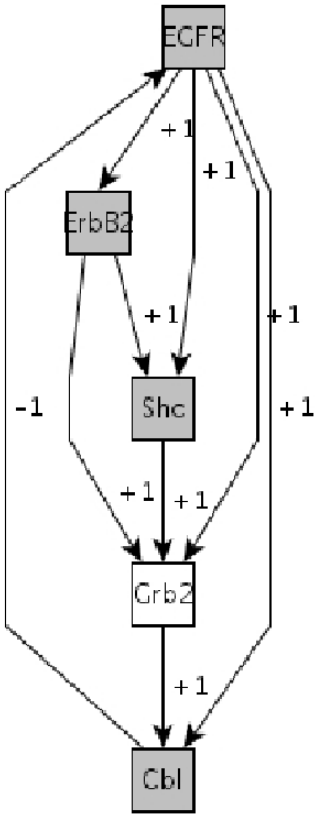

Figura 5.2: *Grafo de um sub-módulo da via EGF. As setas com +1 indicam ativação e -1 repressão. Só o gene em branco é não medido e o gene EGFR é forçado.*

obtidos considerando-se o comportamento linear entre os pontos observados. Geramos os parâmetros com valores, tais que criam-se orbitas interessantes [Fernandez *et al.*], $\alpha_{ji} = 2/\cos(\pi/k)$, onde k é o número de genes no laço e $a_i = \alpha_{ji}/2$.

A partir do grafo na Figura 5.2, testamos a modelagem desta via empregando sistema de equações diferenciais e fizemos as estimativas empregando o algoritmo apresentado anteriormente com dados simulados. Na Figura 5.3, apresentamos o desempenho do algoritmo empregando *simulated annealing*.

Com as estimativas para os parâmetros envolvidos, verificamos que a solução obtida é mínimo local, Figura 5.4.

Na Figura 5.5, verifica-se que o erro entre as duas séries (estimada e real) é muito pequeno.

Nas condições anteriores, o método de estimação foi rodado 30 vezes, obtendo o seguinte resultado: na Figura 5.6, no eixo Y o erro para as estimativas $(\alpha_{estimado} - \alpha_{real})$. Note que, este erro é relativamente pequeno para quase todos os parâmetros em relação ao espaço paramétrico, onde é realizada a busca.

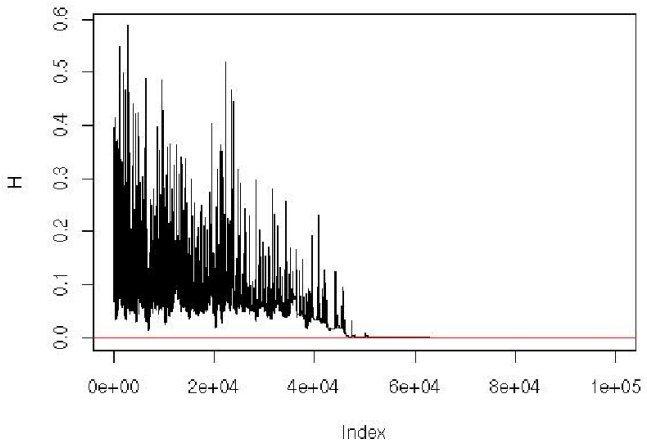

Figura 5.3: *H resultante da estimação com o algoritmo de estimação por simulated annealing, com* $N = 100000$.

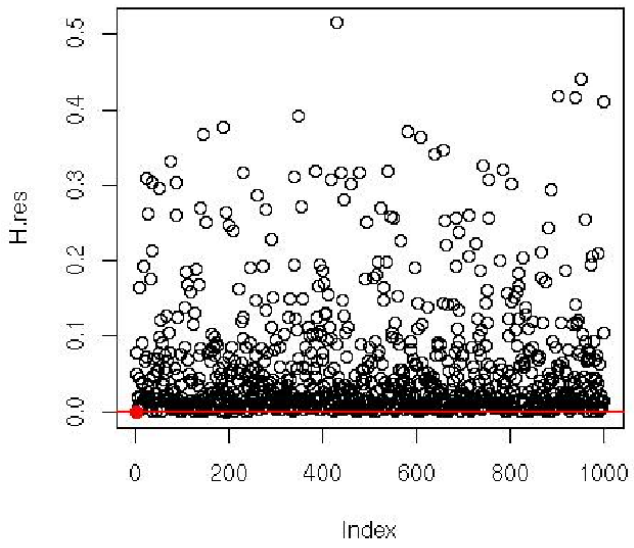

Figura 5.4: *Gerando* $n = 1000$ *perturbações ao redor da estimativa para os parâmetros, confirmamos que a solução obtida é mínimo local. No eixo Y, os valores de H correspondem às perturbações e a linha inferior horizontal indica o menor H (obtido com a estimativa dos parâmetros).*

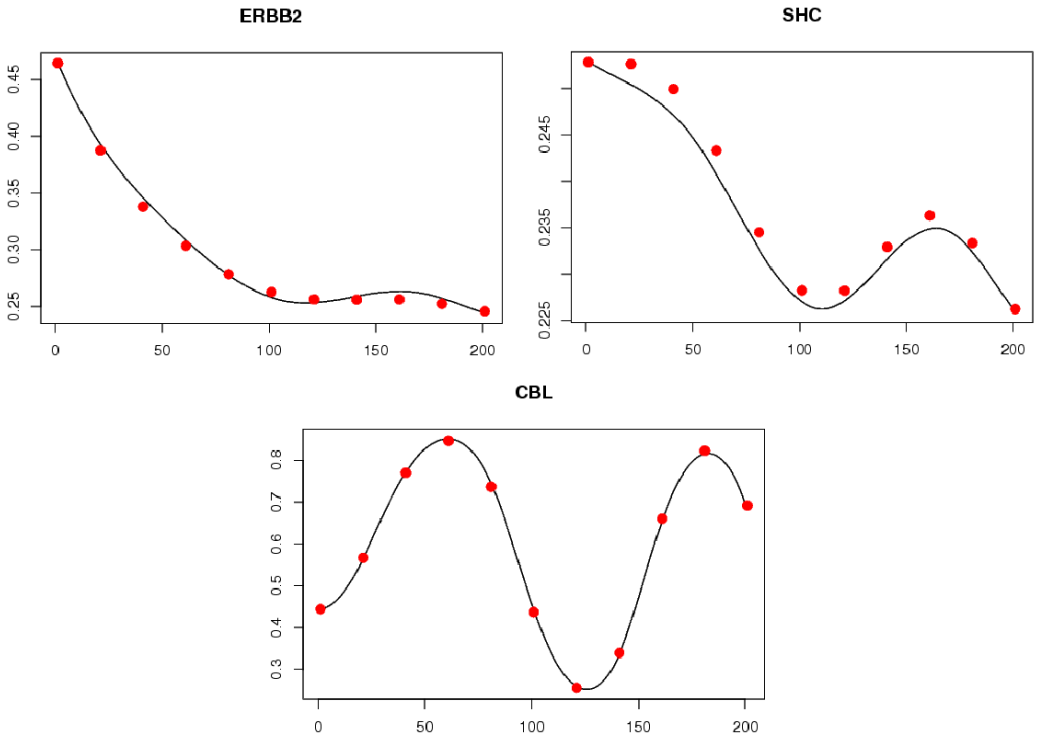

Figura 5.5: *Comparação entre as séries estimadas (curvas) e os valores observados para os genes medidos (pontos). No eixo X, o tempo e no Y, os níveis de expressão gênica.*

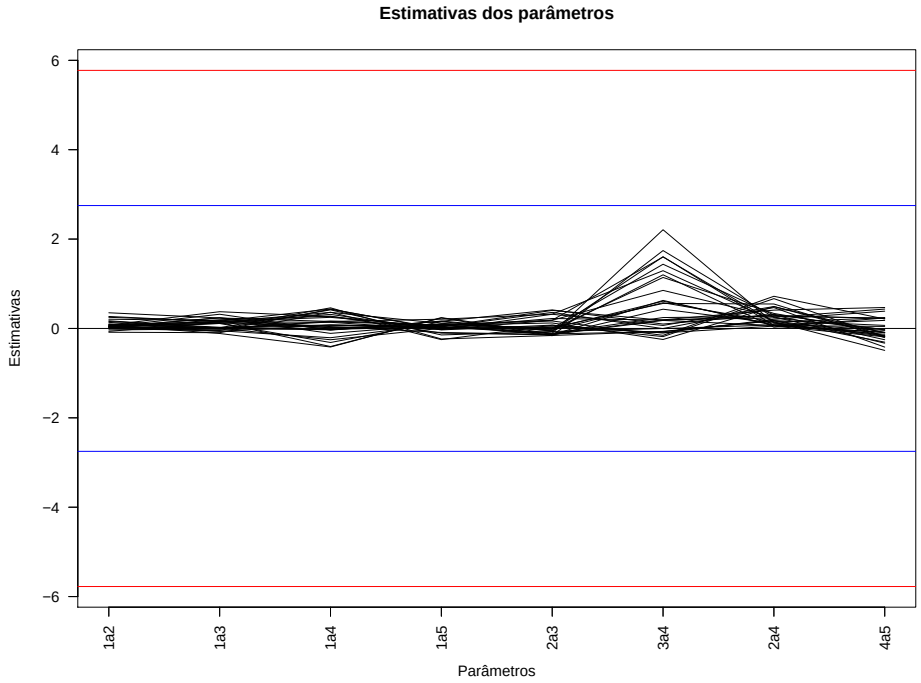

Figura 5.6: *Avaliação da qualidade das estimativas. No eixo X os α's estimados e no eixo Y o erro para as estimativas,* $\alpha_{estimado} - \alpha_{real}$.

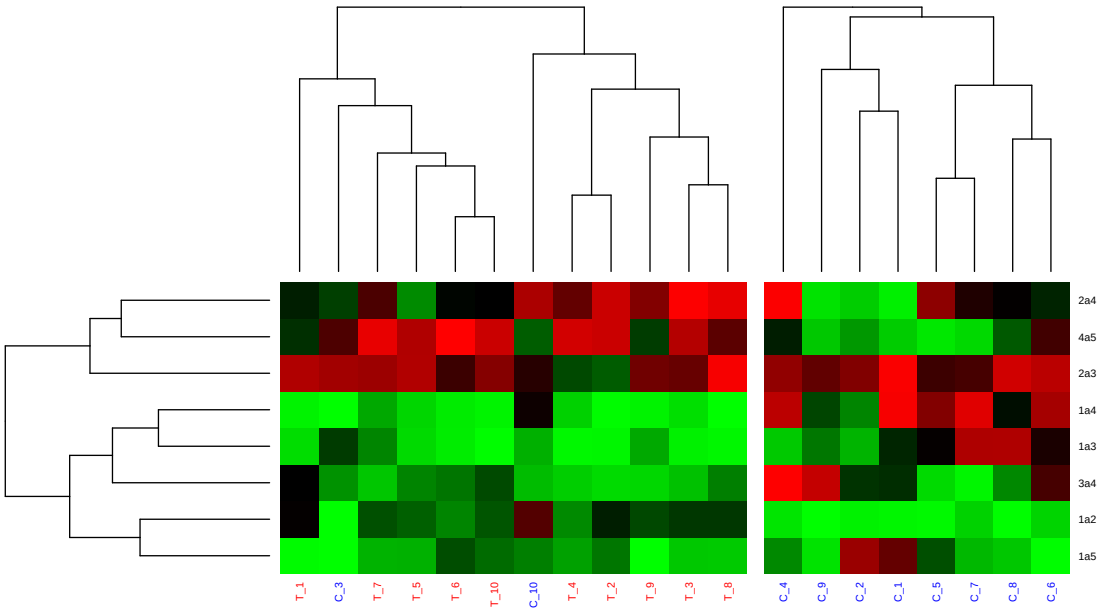

Figura 5.7: *Cluster K-Médias de dois grupos separando Controle e Tratamento em dados reais (Seção* 6.0.1*).*

6
Conjuntos de dados em estudo

6.0.1 Influência de Sulindac em células imortalizadas

Estudo do comportamento de células imortalizadas em duas condições (Tratamento - Sulindac e Controle - DMSO), gerado no Instituto Ludwig no grupo de pesquisa do professor Dr. Luiz Fernando Lima Reis pela pesquisadora Dra. Sibele I. Meireles. Neste conjunto de dados foram observadas duas séries temporais, começando ambas no mesmo ponto inicial ($t = 0$) e sem tratamento, ou seja, controle. Também foram estudados os seguintes instantes de tempo: 3, 6, 12, 24 e 48 horas. Para estes seis instantes de tempo os níves de expressão foram medidos nas duas condições em estudo. Os conjuntos de dados com resultados de séries temporais são muito interessantes para o estudo das redes de regulação gênicas, porque nos permitem avaliar o comportamento das mesmas, pelos níves de expressão, ao longo do tempo.

6.0.2 Resposta à varíola em macacos

Kathleen H. Rubins, Lisa E. Hensley, Peter B. Jahrling, Adeline R. Whitney, Thomas W. Geisbert, John W. Huggins, Art Owen, James W. LeDuc, Patrick O. Brown, and David A. Relman. **"The host response to smallpox: Analysis of the gene expression program in peripheral blood cells in a nonhuman primate model"**. PNAS Vol. 101 No. 42, 15190-15195. October 19, 2004.

Neste conjunto de dados, foi feito um estudo em macacos referente à resposta em varíola. Temos séries temporais de experimento realizados em alguns macacos, submetidos a tratamento, e dos quais foram medidos seus níves de expressão gênica ao longo do tempo.

6.0.3 Identificação de genes expressos no ciclo celular e seu comportamento em tumores

Michael L. Whitfield, Gavin Sherlock, Alok J. Saldanha, John I. Murray, Catherine A. Ball, Karen E. Alexander, John C. Matese, Charles M. Perou, Myra M. Hurt, Patrick O. Brown, and David Botstein. **"Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors"**. Molecular Biology of the Cell Vol. 13, 1977-2000, June 2002.

Neste conjunto de dados, também de séries temporais, temos a evolução de células humanas em vários instantes de tempo. Este conjunto de dados nos resultou de muito interesse, pelo grande número de genes estudados (mais de 16.000). Isto é de grande proveito porque as redes em estudo ficam mais cobertas, ou

seja, mais genes foram medidos, o que torna a análise mais rica e com mais opções.

6.1 Principais Objetivos do Projeto

Os principais objetivos deste projeto são:

① Desenvolver modelos que descrevam de forma simples e clara o comportamento das vias de sinalização com enfoque qualitativo.

② Modelar vias de sinalização, a partir de dados reais a partir de (*microarray*) e simulados, empregando a abordagem de sistema de equações diferenciais (contínua) e redes booleanas (discreta);

③ Ajudar a captar e entender o comportamento das vias de sinalização em condições biológicas diferentes.

④ Avaliar a qualidade dos estimadores relacionada ao erro, consistência e robustez;

⑤ Continuar o desenvolvimento do aluno nesta nova e interdisciplinar área de pesquisa.

Referências Bibliográficas

(1999). Nature genetics supplement. Vol. 21, no. 1.

Albert, R. & Barabási, A. (2002). Statistical mechanics of complex networks. *The American Physical Society* **74**, 47–97.

Brown, K. S. & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Phys Rev E Stat Nonlin Soft Matter Phys* **68**. (2 Pt 1):021904.

Cristo, E. B. (2003). Métodos estatísticos na análise de experimentos de *microarray*.

Dorogovtsev, S. & Mendes, J. (2002). Evolution of networks. *Physics* .

Esteves, G. H. (2007). *Métodos Estatísticos Para a Análise de Dados de cDNA microarray em um Ambiente Computacional Integrado*. Ph.D. thesis, Instituto de Matemática e Estatística da Universidade de São Paulo.

Fernandez, R., Fontes, L. R. & Neves, E. J. (). Density-profile process describing biological signaling networks: Almost sure convergence to deterministic trajectories. January, 2008, submitted.

Gomes, L. I., Esteves, G. H., Carvalho, A. F., Cristo, E. B., Hirata-Jr., R., Martins, W. K., Marques, S. M., Camargo, L. P., Brentani, H., Pelosof, A., Zitron, C., Sallum, R. A., Montagnini, A., Soares, F. A., Neves, E. J. & Reis, L. F. L. (2005). Expression profile of malignant and non-malignant lesions of esophagus and stomach: Differential activity of functional modules related to inflammation and lipid metabolism. *Cancer Research* **65**(16), 7127–7136.

Gomes, L. I., Silva, R. L. A., Stolf, B. S., Cristo, E. B., Hirata-Jr., R., Soares, F. A., Reis, L. F. L., Neves, E. J. & Carvalho, A. F. (2003). Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray. *Analytical Biochemistry* **321**, 244–251.

Gouzé, J. & Sari, T. (). A class of piecewise linear differential equations arising in biological models. INRIA, preprint, http://www.inria.fr/comore.

Haberichter, T., Madge, B., Christopher, R. A., Yoshioka, N., Dhiman, A., Miller, R., Gendelman, R., Aksenov, S. V., Khalil, I. G. & Dowdy, S. F. (2007). A systems biology dynamical model of mammalian G1 cell cycle progression. *Mol. Syst. Biol.* **3**(84).

Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**, S233–S240.

Jong, H. (2002). Modeling and simulation of gene regulatory systems: A literature review. *Journal of Computational Biology* **9**, 67–103.

Jong, H., Gouzé, J., Hernadez, C., Page, M., Sari, T. & Geiselmann, J. (2003). Qualitative simulation of genetic regulatory networks using piecewise linear models. *Bulletin of Mathematical Biology* .

Kim, D., Rath, O., Kolch, W. & Cho, K. H. (2007). A hidden oncogenic positive feedback loop caused by crosstalk between Wnt and ERK pathways. *Oncogene* **26**(31), 4571–4579.

Laub, M., McAdams, H., Feldblyum, T., Fraser, C. & Shapiro, L. (2000). Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290**, 2144–2148.

Legewie, S., Bluthgen, N. & Herzel, H. (2006). Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability. *PLoS Comput. Biol.* **2**(9). :e120.

Meireles, S., Cristo, E., Carvalho, A., Hirata, R., Pelosof, A., Gomes, L., Martins, W., Begnami, M., Zitron, C., Montagnini, A., Soares, F., Neves, E. & Reis, L. (2004). Molecular classifiers for gastric cancers and pre-malignant diseases of the gastric mucosa. *Cancer Research* **64**, 1255–1265.

Poornachandran, R. & Chidambaram, D. (2003). Learning gene networks from microarray data sets. *CSE 591, Computational Molecular Biology* .

Rocke, D. & Durbin, B. (2001). A model for measurement error for gene expression arrays. *Computational Biology* **8**, 557–569.

Schena, M. (2002). *Microarray Analysis*. John Wiley & Sons.

Schneider, D., Latifi, A., Jong, H. & Geiselmann, J. (2003). Dynamics and simulation of the genetic regulatory networks in bacteria. *Recent Research Developments in Genetics* **3**, 55–83.

Smolen, P., Baxter, D. & Byrne, J. (2000). Mathematical modeling of gene networks. *Neuron* **26**, 567.

Stolf, B. S., Carvalho, A. F., Santos, M. M. S., Díaz, J. P., Simão, D. F., Cristo, E. B., Martins, W. K., Esteves, G. E., Hirata-Jr., R., Curado, M. P. & Soares, F. A. (2006). Class distinction between follicular adenomas and follicular carcinomas of the thyroid gland on the basis of their signature expression. *Cancer* **106**(9), 1891–1900.

Thieffry, D. & Romero, D. (1999). The modularity of biological regulatory networks. *BioSystems* **50**, 49–59.

Toni, B., Thieffry, D. & Bulajich, R. (1999). Feedback loops analysis for chaotic dynamics with an application to lorenz system. *Fields Institute Communications* **21**, 473–483.

van Laarhoven, P. J. M. & Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Dordrecht, Holland: D. Reidel Publishing Company.

Weng, G., Bhalla, U. & Iyengar, R. (1999). Complexity in biological signaling systems. *Science* **284**, 92–96.

Zhou, X., Wang, X. & Dougherty, E. (2003). Construction of genomic networks using mutual-information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design. *Signal Processing* **83**, 745–761.