MAC 5701
Tópicos em Ciência da Computação

Tema:
Modelagem de Redes de Expressão Gênica

Aluno: Nestor Walter Trepode
Orientador: Junior Barrera

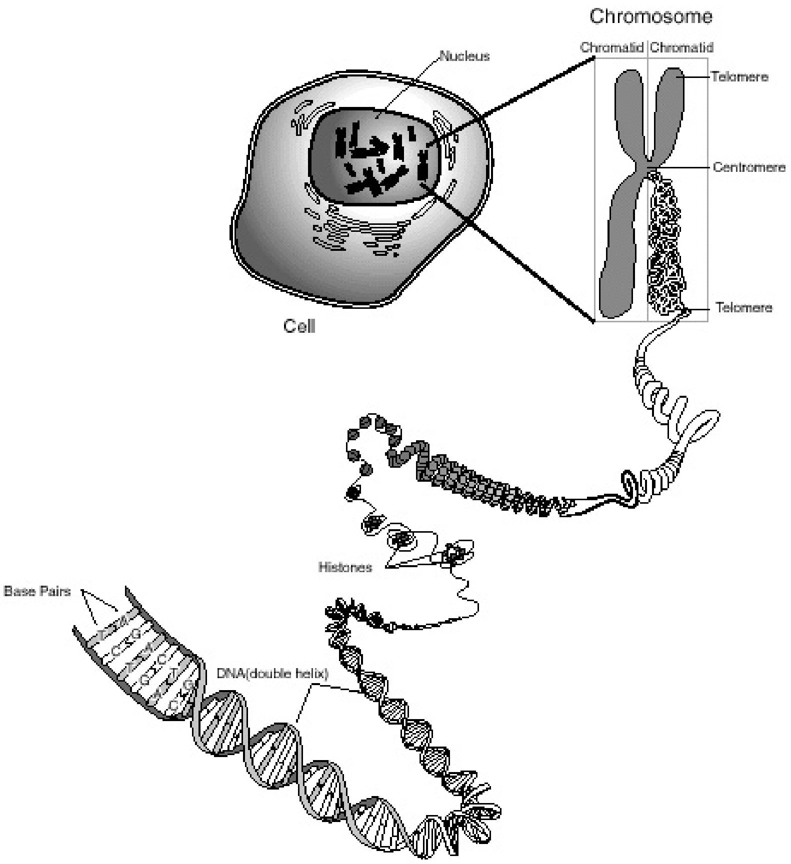


Figura 1: Cromossomo *(fonte: NHGRI, National Human Genome Research Institute).*

1 Área de Estudo

1.1 Redes de Expressão Gênica

Os cromossomos do núcleo da célula são estruturas replicáveis que contem o *DNA –ácido desoxi-ribonucléico–*, molécula que codifica a informação genética. O DNA possui na sua seqüência de nucleotídeos –sub-unidade de DNA ou RNA– um arranjo de genes. Eles estão formados por sub-seqüências ordenadas de nucleotídeos e constituem a unidade física e funcional fundamental da hereditariedade (Figuras 1 e 2). Os genes possuem funções específicas e, para realizar-las, cada gene codifica um produto funcional determinado, i.e. uma proteína ou uma molécula de RNA –ácido ribonucléico–, que é produzido quando ele é expresso. A atividade da célula é determinada por quais genes estão sendo expressos ou não. As *proteínas* são necessárias para a estrutura, função e regulação das células do corpo, tecidos e órgãos, e cada proteína tem funções únicas. Exemplos de proteínas são as hormônios, enzimas e anticorpos.

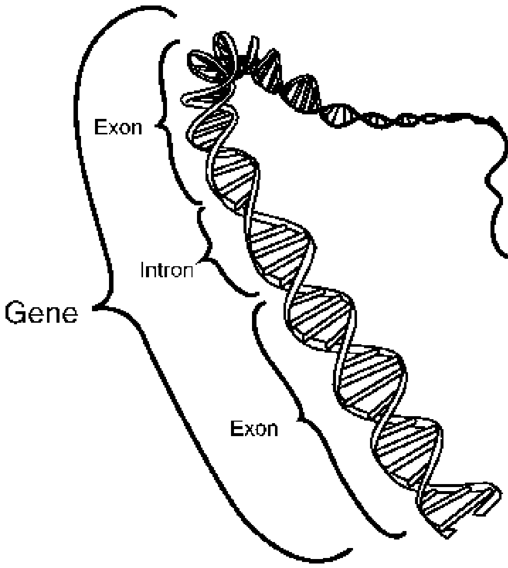


Figura 2: Gene *(fonte: NHGRI, National Human Genome Research Institute).*

A *expressão gênica* é o processo que faz que a informação codificada nos genes seja convertida nas estruturas que estão presentes e atuam dentro da célula. Ela ocorre em dois passos (Figura 3): a) *transcrição* da informação codificada em um gene, mediante a síntese de moléculas de RNA, i.e cópias dele que são enviadas fora do núcleo da célula, e b) *tradução* da informação codificada nos nucleotídeos do mRNA (RNA mensageiro: um dos tipos de RNA produzidos na transcrição), mediante a síntese de uma proteína, a qual está formada por uma seqüência perfeitamente definida de amino ácidos.

O processo da expressão gênica é regulado por um complexo mecanismo de sinais celulares (Figura 4). As proteínas, produzidas como conseqüência da expressão dos genes, formam complexos multi-protéicos inter-atuantes, que interagem a sua vez com sinais externos e, ao mesmo tempo, enviam sinais de realimentação ao núcleo da célula, os quais mudam os padrões de expressão gênica. Os genes e os produtos da expressão gênica formam assim uma rede de transmissão de sinais que regula o funcionamento da célula. Esta é rede chamada de *rede de expressão gênica* (ou de regulação gênica), onde o grau de ativação (ou nível de expressão) de cada gene depende dos níveis de expressão dele mesmo e de outros genes em instantes anteriores, como também de estímulos externos.

1.2 Arquitectura e Dinâmica de uma Rede de Expressão Gênica

Há dois aspectos principais a considerar nestas redes complexas de interações: sua arquitetura e sua dinâmica.

A *arquitetura* de rede consiste na fiação que representa as conexões biomoleculares e as *regras*,

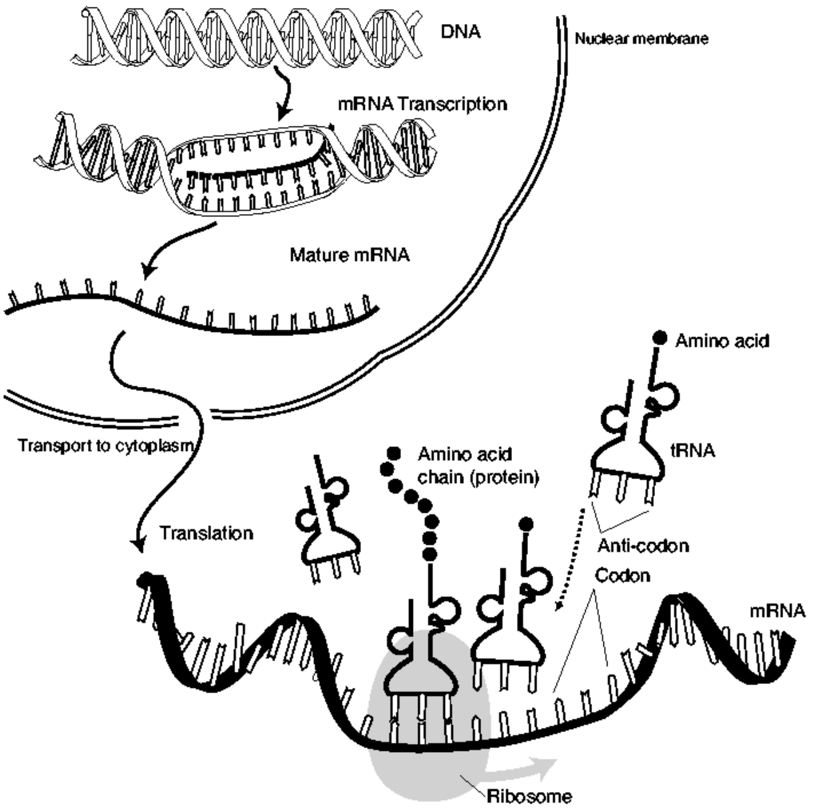


Figura 3: Expressão gênica *(fonte: NHGRI, National Human Genome Research Institute).*

funções ou *códigos* que representam as interações moleculares na célula [1]. Ela define como cada uma das *variáveis* depende das outras. Estas variáveis podem representar valores da atividade molecular, como os níveis de expressão dos genes, moléculas de sinalização, etc., num modelo adotado dado.

A *dinâmica* da rede dá a evolução no tempo das variáveis. O *estado* da rede é o valor destas *variáveis de estado* em um tempo dado. Uma *trajetória* é uma série de transições de estado (por exemplo diferenciação, resposta a perturbações). Quando o sistema retorna a um estado previamente visitado –em um modelo determinístico–. ele ficará seguindo o mesmo ciclo de estados (padrões da expressão). Este ciclo periódico de estados é chamado *atrator* da rede. Os atratores são o resultado final do sistema, por exemplo: fenótipo, tipo da célula, doença crônica, etc.[2].

A atividade da célula é determinada por quais dos seus genes estão sendo expressos ou não. A tecnologia de microarray fornece seus níveis de expressão medindo níveis do mRNA. Estes níveis indicam também a quantidade de proteína disponível na célula em um instante de tempo particular. Considerando múltiplas observações podemos ver a disponibilidade do mRNA no tempo, ou em diferentes condições (por exemplo em resposta aos estímulos, evolução do câncer,

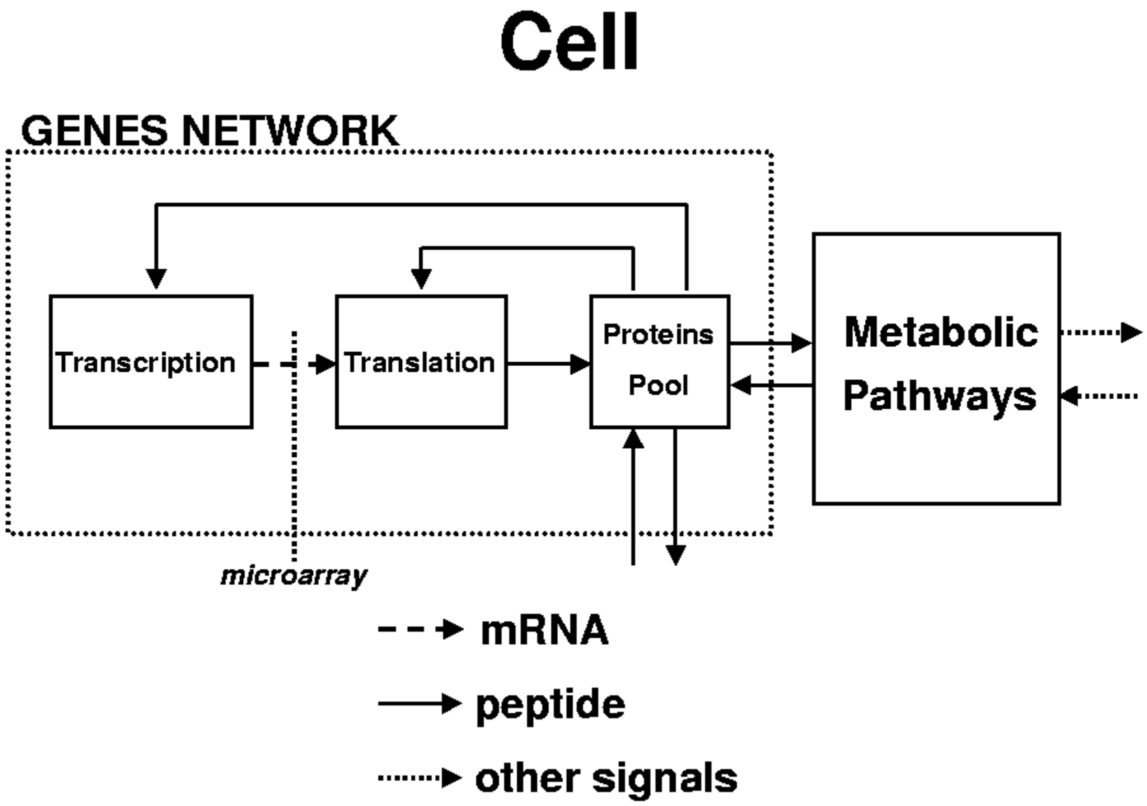


Figura 4: Rede de expressão gênica.

células normais e doentes, etc.). A maioria dos estudos que tentam inferir os mecanismos de expressão na célula usam níveis do mRNA como variáveis de estado porque são as variáveis mais fáceis de medir, em especial por causa destas tecnologias de grande escala de medição da expressão gênica.

As redes de regulação gênica podem ser muito complexas, fazendo que a reconstrução da rede seja um problema difícil. Dados N genes, há um número exponencial de interações possíveis (fiações ou conexões) entre os genes. Este problema de dimensionalidade (ou insuficiência de dados para o tamanho do problema) faz a tarefa da modelagem, identificação e simulação da rede mais difícil. A seleção cuidadosa das variáveis da entrada, sem perder nenhuma das importantes, e o uso de informação previa sobre o que é sabido ou possível biologicamente, podem ser cruciais para tratar com o problema da dimensionalidade.

1.3 Modelos de Redes Genéticas

Vários tipos de modelos de redes de regulação gênica tem sido propostos. Entre as opções disponíveis temos:

- Determinísticas ou Estocásticas
- Discretas ou Continuas

Uma rede *determinística* é um sistema rígido onde o estado da expressão gênica (nível da expressão de todos os genes que estão sendo considerados) em um tempo dado e as interações regulatórias entre eles determinam sem ambigüidades o estado de expressão gênica no tempo seguinte. Em tal rede, há somente um trajeto possível de um estado da expressão do gene a outro [5].

Em um sistema *estocástico*, um estado dado de expressão gênica pode gerar mais de um estado seguinte, por exemplo, fazendo que diferentes células da mesma população podem seguir um trajeto diferente de expressão gênica de um estado para o próximo. A estocacidade permite variações significativas na seqüência da ativação e inativação dos genes.

A natureza estocástica de redes genéticas é suportada por considerações teóricas e por resultados experimentais. Os mecanismos que explicam esta estocacidade são, por exemplo, a degradação dos produtos do gene, a colisão espacial necessária antes que um reagente possa exercer sua influência, equações de reação reversível, etc. [3].

Conseqüentemente, os modelos estocásticos descrevem a cinética da regulação gênica mais exatamente do que uma aproximação determinística. Em geral, os resultados experimentais podem ser melhor explicados e modelados por mecanismos estocásticos. Em compensação, os modelos estocásticos são mais complexos do que os determinísticos.

Uma rede *Booleana* é a rede discreta mais simples. Consiste de n nós, representando os genes, que podem estar expressos ou não (estado 1 ou 0, respectivamente). A dinâmica da rede é determinada pelas n funções booleanas (uma para cada nó), que recebem k nós de entrada e determinam o estado seguinte para esse nó a partir dos estados de todos os nós de entrada [6, 7].

São, naturalmente, uma simplificação grosseira de uma rede genética porque a expressão gênica tende a ser contínua (ou, ao menos, multi-valuada) antes que binária —nunca é um caso de todo-ou-nada— e as relações possuem valores contínuos (por exemplo, a concentração das moléculas varia suavemente).

Não obstante, as redes booleanas dão uma estrutura em que os genes podem ter interações complexas e mostrar o comportamento comparável às características das redes genéticas biológicas (por exemplo comportamento complexo global, auto-organização, etc. [8]). Podem ser um bom ponto de partida para uma modelagem realística das redes genéticas [15].

Em uma *Rede Booleana Aleatória* (RBN: "Ramdom Boolean Network"), cada nó – gene – pode ter um esquema diferente de entrada/saída, e um número diferente k das entradas. Cada nó pode também ter diferentes funções booleanas escolhidas aleatóriamente. São um refinamento probabilístico mais realístico das redes booleanas [13, 14].

Uma Rede Genética pode também ser modelada como um conjunto de *equações diferenciais* não lineares [10]. Devem ser encontrados os parâmetros, que indicam a taxa de mudança de cada gene. Eles são: $n \times n$ pesos de interação gênica, n termos de polarização e constantes de tempo para cada um dos n nós do sistema. A suposição de passos de tempo discretos para o próximo estado da rede, é aqui desnecessária. Chen et o al. (1999) [11] conseguiram reduzir o espaço da busca dos parâmetros fazendo algumas suposições adicionais. Desta maneira, eles podem considerar mais características do que os modelos mais simples em seu modelo transcripcional linear.

1.4 Identificação da Rede

Em *identificação de redes* o objetivo é construir um modelo da rede de interações regulatórias entre os genes. Isto requer a inferência das relações causais entre genes, isto é. encontrar a arquitetura de rede a partir dos seus padrões de expressão.

Para um conjunto dado de medidas, os métodos de *engenharia reversa* tentam inferir a rede regulatória subjacente desconhecida. Usam geralmente um modelo paramétrico, e ajustam os parâmetros aos dados reais.

Se a arquitetura da rede regulatória for desconhecida, o modelo paramétrico terá que ser muito geral e simplista. Se a arquitetura fosse bem conhecida, nós poderíamos usar um modelo mais detalhado estimar os parâmetros relacionados aos mecanismos individuais [4]. A identificação prévia da arquitetura [9] serve para restringir a conectividade da rede. Isto pode reduzir dramaticamente a quantidade de dados amostrais necessários para uma estimação boa dos parâmetros.

Diversos métodos para a identificação de redes genéticas – a arquitetura e a dinâmica – foram propostos. Alguns deles são listados abaixo.

Somogyi et al [9] propuseram duas premissas para a identificação de redes genéticas. A primeira (enfoque reducionista-mecanicista) é o seguinte:

Premissa: "Um gene para cada função e uma função para cada gene"

- Redução completa do organismo a genes
- Determinação das estruturas e atividades das proteínas
- Mapeamento das interações dos produtos genéticos moleculares
- Montagem em banco de dados dos mecanismos moleculares
- Síntese do modelo computacional como a soma de suas partes

A segunda (heurística, estratégia de sistemas integrativos) é:

Premissa: "A função dos genes esta distribuída numa rede de processamento paralelo"

- Identificar os genes e os elementos da rede genética
- Determinar os estados da rede (padrões de expressão)
- Mapear trajetórias alternativas e atratores
- Trajetórias paralelas sugerem entradas compartilhadas
- Ondas de perturbação confinadas determinam ligações temporais

- Realizar engenharia reversa computacional da rede

O algoritmo de Akutsu et al. [7] pode identificar uma rede booleana de n nós a partir de $O[\log n]$ pares da transições de estado, procurando exaustivamente todas as funções booleanas possíveis até que um conjunto que coincida com todos os dados seja encontrado.

O algoritmo da engenharia reversa (*REVEAL: REVerse Engineering ALgorithm*) [6] é baseado na comparação das entropias de Shannon dos dados da entrada e da saída que dão o mínimo grau de entrada k para cada nó, e assim a rede efetiva mínima pode ser inferida. Depois, eles fazem uma busca exaustiva para a regra que se ajusta os dados, como o algoritmo de Akutsu et al..

O algoritmo *Predictor* [12] determina o conjunto de redes booleanas consistentes com um conjunto observado de perfis de expressão gênica de estado estacionário, cada um gerado por uma perturbação diferente da rede genética. Este método é usado iterativa e interativamente com o método *Chooser*, que usa um enfoque baseado em entropia para propor um experimento de perturbação adicional para discriminar entre conjuntos de redes determinadas pelo predictor. Nesta maneira a rede genética é refinada sucessivamente usando os dados cumulativos de expressão.

Wahde e Hertz [10] modelaram redes de regulação gênica no contexto de redes neurais recorrentes de tempo contínuo e propuseram um método para determinar os parâmetros de tais redes.

Nos propusemos (Barrera et al., 2002) [16] uma técnica para estimar a dinâmica de redes multi-valoradas discretas a partir de dados de trajetórias e sistemas de envelope (isto é, um par de sistemas com trajetórias acima e abaixo do sistema de alvo) obtidos por simulações exaustivas, baseadas em conhecimento biológico sobre a rede. O envelope restringe o espaço de sistemas candidatos e simplifica assim o problema de estimação.

O nosso método utiliza o modelo matemático dos *sistemas dinâmicos finitos* ou FDS ("Finite Dinamical Systems") que são adequados para modelar uma rede de expressão gênica. Neste modelo o nível de expressão de cada gene é representado pelo valor de uma *variável de estado*. O *estado* de um FDS é definido pelo valor num instante dado de todas as variáveis de estado. O valor de uma variável de estado num instante de tempo depende dos estados anteriores e de estímulos externos anteriores através de uma função que representa os mecanismos de regulação gênica descritos anteriormente. Cada uma destas funções é uma componente de um vetor de funções chamado de *função de transição*, a qual define a transição de um estado do FDS para o seguinte.

A *identificação* de um sistema dinâmico finito consiste em encontrar as componentes da função de transição a partir de amostras dos estados do sistema. Enquanto a medição de muitas variáveis pode permitir a obtenção de um modelo mais exato, a complexidade do modelo (e a dificuldade para acha-lo) aumenta exponencialmente com o número de variáveis. A teoria de Aprendizado Computacional (que se aplica para a identificação) explica que quanto maior seja o número de variáveis necessárias, mais difícil é a tarefa de modelagem, porque o tamanho do espaço de busca (número de modelos possíveis) aumenta exponencialmente com o número de parâmetros do modelo. A diminuição do tamanho do espaço dos modelos plausíveis mediante a aplicação de restrições adicionais pode simplificar consideravelmente a busca do melhor modelo. A restrição do modelo pelo uso de informação prévia sobre o que é biologicamente conhecido ou possível é, provavelmente, uma das mais importantes estratégias de que se dispõe para lidar com o problema da dimensionalidade. A precisão com que esta informação seja incluída no processo de identificação dependerá da forma em que é feita a modelagem desta informação disponível.

2 Modelagem e Simulação de Redes de Expressão Gênica

Foram estudados diferentes modelos matemáticos atualmente sendo utilizados para a representação, simulação e identificação de redes de regulação gênica. Foram analisados modelos discretos e contínuos, determinísticos e estocásticos.

Os tópicos cobertos neste estudo são

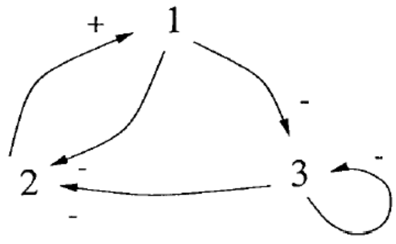
- Grafos Dirigidos e não Dirigidos
- Redes Bayesianas
- Redes Booleanas
- Redes Lógicas Generalizadas
- Equações Diferenciais Ordinárias não Lineares
- Equações Diferenciais Lineares por Partes
- Equações Estocásticas Principais
- Cadeias de Markov
- Aprendizado de Cadeias de Markov

O material estudado será apresentado em seminários e estão sendo desenvolvidos experimentos de modelagem e simulação de redes de expressão gênica mediante o uso de cadeias de Markov.

2.1 Grafos Dirigidos e não Dirigidos

È talvez a forma mais simples de modelar uma rede de regulação gênica. Um grafo dirigido G é definido como uma tupla $< V, E >$, onde V é um conjunto de vértices e E é um conjunto de arestas. Uma aresta dirigida é uma tupla $< i, j >$ de vértices, onde i é o fim e j é o início da aresta. Os vértices do grafo dirigido representam genes ou outros elementos de interesse no sistema de regulação, enquanto que as arestas denotam interações entre os genes. Esta representação por grafos da rede de regulação pode ser generalizada de varias maneiras. Os vértices e arestas podem ser rotulados, por exemplo, para expressar informações dos genes e suas interações. Definindo uma aresta dirigida como uma tupla $< i, j, s >$, como s igual a $+$ ou $-$, pode se indicar se i é ativado ou inibido por j (Figura 5 (a) e (b)). Podem usar-se hipergrafos para representar situações nas quais varias proteínas conjuntamente regulam a expressão de um gene. As arestas são agora definidas por $< i, J, S >$, onde J representa uma lista de genes reguladores e S uma lista correspondente que indica sua influência regulatória (Figura 5 (c) e (d)).

As atuais bases de dados e de conhecimento com informações de interações regulatórias podem ser vistas em grande parte como representações de grafos com muitas anotações [18, 19,

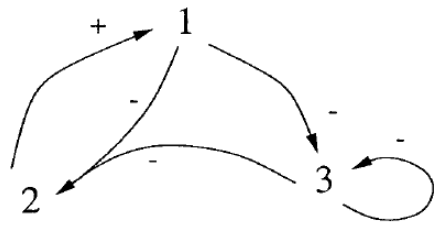


(a)

$$V = \{1, 2, 3\}$$

$$E = \{\langle 2, 1, - \rangle, \langle 3, 1, - \rangle, \langle 1, 2, + \rangle, \langle 2, 3, - \rangle, \langle 3, 3, - \rangle\}$$

(b)



(c)

$$V = \{1, 2, 3\}$$

$$E = \{\langle 2, [1, 3], [-, -] \rangle, \langle 3, [1], [-] \rangle, \langle 1, [2], [+] \rangle, \langle 3, [3], [-] \rangle\}$$

(d)

Figura 5: (a) Representação de uma rede de regulação gênica mediante um grafo dirigido e (b) sua definição. (c) (d) Representação de uma rede de regulação gênica com interações cooperativas mediante um hiper-grafo dirigido.

20, 21, 22, 23, 24, 25]. Usualmente, estão complementadas com aplicativos para compor e editar redes, selecionando e manipulando as interações individuais. Como mínimo, permitem ao usuário visualizar redes completas ou parciais, muitas vezes com diferentes níveis de detalhe, e navegar as redes [26, 27]. Algumas bases de dados e conhecimento atuais integram vias metabólicas e de transmissão de sinais com processos de regulação gênica [19, 23].

Diversas operações de grafos podem realizar-se para fazer predições biológicas relevantes acerca dos sistemas regulatórios. Uma busca de caminhos entre dois genes, por exemplo, pode revelar interações regulatórias faltantes o dar pistas de redundância na rede. Além disso, ciclos na rede indicam realimentações que são importantes na homeostase[1] e diferenciação. As características de conectividade global da rede, tais como a média e a distribuição do número de reguladores por gene, são uma indicação da complexidade da rede. Sub-grafos fracamente conectados indicam módulos funcionais do sistema regulatório cujo comportamento pode ser considerado isoladamente. A partir da comparação de redes regulatórias de diferentes organismos, pode-se estabelecer até que ponto partes do sistema regulatório tem sido evolutivamente conservadas.

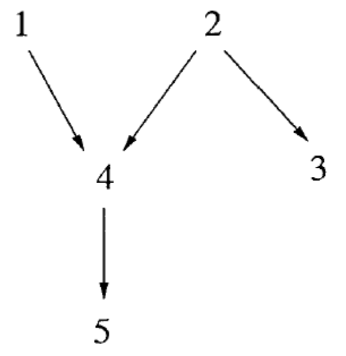
Os modelos de grafos aos quais se aplicam as operações anteriores podem ser construídos a partir de informações armazenadas em bases de dados o de conhecimento, mas também podem

[1] A tendência de uma população a alcançar um ponto de equilíbrio genético e resistir mudanças

obtidas de dados de expressão gênica por enfoques *indutivos* ou de *engenharia reversa*. Uma grande variedade de algoritmos de *clustering* tem sido usados para agrupar juntos genes com padrões de expressão temporal semelhantes. O uso de algoritmos de clustering é motivado pela idéia de que dois genes que apresentam padrões de expressão semelhantes no tempo podem se regular um ao outro ou serem regulados por um terceiro gene. Analises adicionais podem permitir a extração de mais informação de supostas conexões regulatórias entre genes co-expressos no grafo, tais como a analise de atrasos de tempo [28, 29, 30], experimentos de resposta a perturbações [31, 32, 33, 34], mapeamento de genes co-expressos em funções de trajetória [23, 35], etc.

2.2 Redes Bayesianas

As redes bayesianas modelam a estrutura de um sistema de regulação gênica mediante um grafo acíclico dirigido $G =< V, E >$ (Fig. 4). Os vértices $i \in V$, $1 \leq i \leq n$, representam genes ou outros elementos e correspondem a variáveis aleatórias X_i. Se i é um gene, então X_i descreve o nível da expressão de i. Para cada X_i, é definida uma distribuição condicional $p(X_i|pais(X_i))$, onde $pais(X_i)$ denota as variáveis que correspondem aos reguladores diretos de i em G. O grafo G e as distribuições condicionais $p(X_i|pais(X_i))$, que definem a rede bayesiana, especificam unicamente uma distribuição conjunta de probabilidade $p(\mathbf{X})$.



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

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

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

$i(X_1; X_2, X_3),\ i(X_2; X_1),\ i(X_4; X_3 \mid X_1, X_2)$

$i(X_3; X_1, X_4, X_5 \mid X_2),\ i(X_5; X_1, X_2, X_3 \mid X_4)$

Figura 6: Exemplo de Rede Bayesiana que consiste de um grafo, distribuições de probabilidade condicional para as variáveis aleatórias, a distribuição de probabilidade conjunta, e as independências condicionais [36].

A *independência condicional* $i(X_i; \mathbf{Y}|\mathbf{Z})$ expressa o fato que X_i é independente de $\mathbf{Y}$ dado $\mathbf{Z}$, onde $\mathbf{Y}$ e $\mathbf{Z}$ denotam conjuntos de variáveis. O grafo codifica a *suposição de Markov*, a qual indica que para cada gene i em G, $i(X_i; nao - descendentes(X_i)|pais(X_i))$ Por meio da suposição de Markov, a distribuição de probabilidade conjunta pode ser decomposta em

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i|pais(X_i)). \tag{1}$$

O grafo implica independências condicionais adicionais, como é mostrado na Figura 6 para a rede do exemplo. Dois grafos, e portanto duas redes bayesianas, são ditas de equivalentes, se implicarem o mesmo conjunto de independências. Os grafos numa classe de equivalência não podem ser distinguidos pela observação em $\mathbf{X}$. Grafos equivalentes podem ser caracterizados formalmente como tendo o mesmo grafo não dirigido subjacente, mas podem discordar no sentido de algumas arestas [36].

Dado um conjunto de dados de expressão D na forma de valores independentes para $\mathbf{X}$, as técnicas de aprendizado de redes bayesianas [37] permitem inferir a rede, ou melhor a classe de equivalência de redes, que melhor se ajusta ao conjunto de dados D. Basicamente, as técnicas atribuem uma pontuação ao grão de casamento ("matching score") para avaliar as redes com respeito aos dados e procurar pela rede com a maior pontuação. Como é sabido que este problema de otimização é NP-complexo, devem ser usados métodos de busca heurísticos que não garantem uma solução global ótima. Um problema adicional, é que os dados de expressão atualmente disponíveis não permitem determinar a rede, dado que no melhor dos casos umas poucas dezenas de experimentos fornecem informação do nível de transcrição de milhares de genes.

Friedman e seus colegas [36] propuseram um algoritmo heurístico para a indução de redes bayesianas a partir de dados de expressão que pode lidar com este problema de insuficiência de dados. Em vez de procurar uma única rede, ou uma única classe de equivalência de redes, focalizam em características que são comuns às redes de elevada pontuação. Em particular, procuram relações de Markov e outras relações entre pares de variáveis X_i e X_j. Uma relação de Markov existe, se X_i é parte do conjunto de variáveis minimal protege X_j do resto das variáveis. Uma relação da ordem existe, se X_i for um pai de X_j em todos os grafos de uma classe de equivalência. Uma relação da ordem entre duas variáveis pode indicar uma relação causal entre os genes correspondentes. Critérios estatísticos para avaliar a confiança nas características foram desenvolvidos. Uma extensão recente do método [38] pode tratar com mutações genéticas e considera características adicionais, como a ativação, a inibição, e relações entre variáveis.

O enfoque de rede bayesiana para a modelagem de redes regulatórias é atrativo pela sua sólida base em estatística, o que permite tratar com aspectos estocásticos da expressão gênica e das medições ruidosas de uma maneira natural. Além disso, as redes bayesianas podem ser usadas quando somente se dispõe de conhecimento incompleto sobre o sistema. Embora as redes bayesianas e os modelos de grafos vistos anteriormente sejam representações intuitivas das redes de regulação gênica, têm a desvantagem de deixar implícitos aspectos dinâmicos da regulação gênica. Em alguma medida, isto pode ser solucionado com generalizações como as redes bayesianas dinâmicas, que permitem modelar relações de realimentação entre genes [39].

2.3 Redes Booleanas

Como una primeira aproximação, o estado de um gene pode ser descrito como uma variável booleana que expressa se ele está ativo (1) ou inativo (0) e em conseqüência se os produtos de sua expressão estão presentes ou ausentes. Além disso, as interações entre elementos podem ser representadas por funções Booleanas que calculam o estado de um gene a partir da ativação de outros genes.

O vetor n-dimensional $\hat{\mathbf{x}}$ de variáveis de uma rede Booleana representa o estado do sistema regulatório de n elementos. Cada $\hat{x}_i$ tem o valor 1 ou 0, de modo que o espaço de estados do sistema consiste de 2^n estados. O estado $\hat{x}_i$ de um elemento no tempo $t+1$ é computado por meio de uma função Booleana ou regra $\hat{b}_i$ a partir do estado de k dos n elementos no tempo precedente t (Notar que k pode ser diferente para cada $\hat{x}_i$). A variável $\hat{x}_i$ é chamada também de *saída* do elemento e as k variáveis a partir do qual ele é calculado de *entradas*. Para k entradas, o número total de possíveis funções Booleanas $\hat{b}_i$ que mapeam as entradas para a saída é 2^{2^k}. Resumindo, a dinâmica de uma rede Booleana que descreve um sistema regulatório é dada por

$$\hat{x}_i(t+1) = \hat{b}_i(\hat{\mathbf{x}}(t)), \quad 1 \leq i \leq n, \tag{2}$$

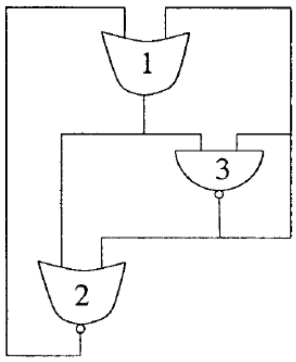
onde $\hat{b}_i$ mapeia k entradas para um valor de saída.

A estrutura de uma rede booleana pode ser re-arranjada na forma de um diagrama de conexões (Figura 7(c)). A fila da parte superior alista o estado no tempo t e a da parte inferior o estado no tempo $t+1$. A função Booleana que calcula a saída a partir das entradas se mostra abaixo de cada elemento. O diagrama de conexões é uma representação conveniente para o cálculo de transições entre estados. Por exemplo, dado um vetor de estados 000 em $t = 0$, o sistema do exemplo mover-se-á para o estado 011 no próximo instante de tempo $t = 1$ Isto é, se os três genes forem inativos em $t = 0$, o segundo e terceiro gene tornar-se-á ativo no instante de tempo seguinte. Assim, as transições entre estados são *determinísticas*, com um único estado de saída para um estado de entrada dado, e *sincrônico*, no sentido de que as saídas dos elementos são atualizadas simultaneamente.

Uma seqüência de estados conectados por transições forma uma *trajetória* do sistema. Dado que o número de estados no espaço de estados é finito, o número de estados numa trajetória será finito também. Mais especificamente, todos os estados iniciais de uma trajetória eventualmente alcançarão um estado estacionário ou um ciclo de estados, chamados também *atrator pontual* ou *atrator dinâmico*, respectivamente. Os estados que não formam parte de um atrator são chamados de estados *transitórios*. Os estados de um atrator e os estados transitórios que levam ao atrator, todos em conjunto constiúem a *bacia de atração* ("basin of attraction").

Com a exceção de redes muito simples, são requeridos programas de computador para o cálculo dos atratores e suas bacias de atração. Exemplos de tais programas são o DDLab, desenvolvido por Andy Wuensche (1998), e o Gene-O-Matic [40], que permite simular a diferenciação em sistemas multicelulares [41].

As redes Booleanas estão entre os primeiros formalismos para os quais foram propostos métodos de inferência do modelo. Um exemplo é o algoritmo REVEAL [6] (outros exemplos e idéias seminais em [7, 12, 42, 43, 44, 45, 46]). Este algoritmo a teoria da informação para estabelecer quantos elementos dados estão conectados na rede e depois determina, a partir dos



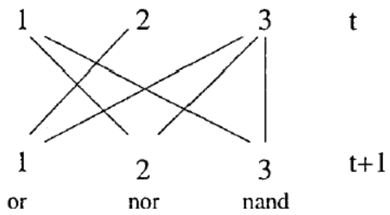
(a)

$$\hat{x}_1(t+1) = \hat{x}_2(t) \text{ or } \hat{x}_3(t)$$

$$\hat{x}_2(t+1) = \hat{x}_1(t) \text{ nor } \hat{x}_3(t)$$

$$\hat{x}_3(t+1) = \hat{x}_1(t) \text{ nand } \hat{x}_3(t)$$

(b)



(c)

Figura 7: (a) Exemplo de rede Booleana (b) Equações correspondentes. Neste caso $n = 3$ y $k = 2$. (c) Diagrama de conexões da rede Booleana.

dados, as funções que especificam a lógica das interações. Para cada um dos n elementos, são consideradas todas as possíveis combinações de k elementos $1 \leq k \leq n$), até que é encontrado um conjunto de entradas que determina totalmente a saída do elemento. Em termos da teoria da informação, isto significa que a *informação mútua* da saída do elemento e suas entradas iguala o valor de informação da saída do elemento. Depois, é encontrada a função para o elemento comparando as transições de estado da trajetória observadas com as definições das funções Booleanas.

Exemplos de modelos de redes Booleanas para sistemas regulatórios simples podem ser encontrados em [47]. Uma aplicação interessante das redes Booleanas é seu uso no estudo de propriedades globais de sistemas regulatórios de grande escala [8, 48, 49, 50, 51]. A idéia básica é gerar redes Booleanas aleatórias com propriedades locais que valem para todos os membros de uma classe de sistemas. Propriedades interessantes são, por exemplo, o número k de reguladores de um gene e o tipo de funções $\hat{b}_i$ mediante as quais os reguladores influenciam a expressão gênica. Localizando os atratores, trajetórias, e bacias de atração no espaço de estados, é possível investigar sistematicamente as implicações das propriedades locais para a dinâmica global da rede.

As redes booleanas permitem analisar de maneira eficiente um grande número de redes regulatórias, fazendo fortes suposições simplificadoras sobre a estrutura e a dinâmica de um sistema de regulação gênica. No formalismo das redes Booleanas, um gene é considerado como ativo ou inativo e são desprezados os níveis intermediários de expressão. Também, é suposto que as transições entre os estados de ativação dos genes ocorrem sincrônicamente. Quando as transições não acontecem simultaneamente, como é usualmente o caso, alguns comportamentos podem não ser preditos pelo algoritmo de simulação. Há situações nas quais as idealizações implícitas no formalismo das redes Booleanas não são apropriadas e métodos mais gerais são requeridos.

2.4 Redes Lógicas Generalizadas

O método baseia-se em um formalismo que generaliza as redes Booleanas permitindo que as variáveis tenham mais de dois valores e que as transições ocorram asincrônicamente [52, 53, 54].

O formalismo usa variáveis discretas $\hat{x}_i$, chamadas *variáveis lógicas*, que são abstrações das variáveis de concentrações reais x_i. Os possíveis valores de $\hat{x}_i$ são definidos comparando as concentrações x_i com os limiares da influência de i sobre outros elementos do sistema regulatório.

Se um elemento i influencia p outros elementos do sistema regulatório, ele pode ter até p limiares diferentes [55]

$$\sigma_i^{(1)} < \sigma_i^{(2)} < \ldots < \sigma_i^{(p)}$$

Dados estes limiares, $\hat{x}_i$ tem os possíveis valores $0, \ldots, p$ definidos da seguinte maneira:

$$\begin{array}{rcl} \hat{x}_i & = & 0, \text{ se } x_i < \sigma_i^{(1)} \\ \hat{x}_i & = & 1, \text{ se } \sigma_i^{(1)} < x_i < \sigma_i^{(2)} \\ & \cdots & \\ \hat{x}_i & = & p, \text{ se } x_i > \sigma_i^{(p)} \end{array}$$

O vetor $\hat{\mathbf{x}}$ denota o estado lógico do sistema regulatório.

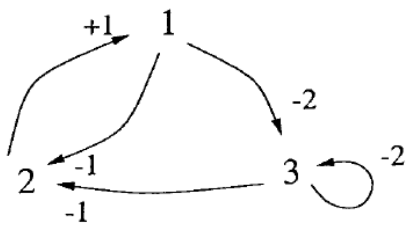
O padrão de interações regulatórias no sistema é descrito por equações lógicas da forma

$$\hat{X}_i(t) = \hat{b}_i(\hat{\mathbf{x}}(t)), \quad 1 \leq i \leq n, \tag{3}$$

onde $\hat{X}_i$ é chamado a *imagem* de $\hat{x}_i$. A imagem é o valor para o qual $\hat{x}_i$ tende quando o estado lógico do sistema é $\hat{\mathbf{x}}$. A imagem de $\hat{x}_i$ não é necessariamente seu valor sucessor, e deve portanto distinguir-se ele de $\hat{x}(t+1)$ em (2). A função lógica $\hat{b}_i$ é uma generalização das funções Booleanas em (2), dado que as variáveis lógicas tem agora mais de dois valores possíveis. A função lógica calcula a imagem de $\hat{x}_i$ a partir do estado lógico do sistema, mais especificamente dos valores de k de seus n elementos.

A Figura 8 mostra um exemplo de rede regulatória. O gene 1 regula os genes 2 e 3, de modo que ele tem dois limiares e a correspondente variável lógica $\hat{x}_1$ toma seu valor de $\{0, 1, 2\}$. Similarmente, $\hat{x}_2$ e $\hat{x}_3$ tem um e dois limiares, respectivamente, e portanto valores possíveis $\{0, 1\}$ e $\{0, 1, 2\}$. Cada aresta do grafo é rotulada com o número de limiar e o signo de sua influência. Por exemplo, o rótulo -2 do gene 1 para o gene 3 significa que 1 inibe 3 acima do seu segundo limiar, isto é, quando $\hat{x}_1 = 2$.

O grafo da Figura 8(a) pode ser transformado em equações lógicas, como se mostra na Figura 8(b). As funções $\hat{b}_1, \ldots, \hat{b}_3$ devem ser especificadas de modo que sejam consistentes com as restrições referentes aos limiares no grafo. A Figura 8(c) mostra exemplos de funções lógicas permitidas pelo método, usando uma notação que enfatiza sua correspondência com as funções Booleanas vistas anteriormente. Consideremos o caso de $\hat{b}_2$. Se $\hat{x}_1 > 0$ e $\hat{x}_3 > 0$, de modo que x_1 e x_3 tenham valores acima do seu primeiro limiar, as influências inibidoras dos genes 1 e 3 sobre o gene 2 se tornam operativas. A Figura 8(c) indica que $\hat{x}_2$ tenderá a 0, ou seja, abaixo do primeiro limiar da proteína produzida pelo gene 2. Se $\hat{x}_1 = 0$ ou $\hat{x}_3 = 0$, isto é, se



(a)

$$\hat{X}_1 = \hat{b}_1(\hat{x}_2)$$
$$\hat{X}_2 = \hat{b}_2(\hat{x}_1, \hat{x}_3)$$
$$\hat{X}_3 = \hat{b}_3(\hat{x}_1, \hat{x}_3)$$

(b)

$\hat{b}_1(\hat{x}_2)$	$\hat{x}_2$
0	0
2	1

		$\hat{x}_3$		
	$\hat{b}_2(\hat{x}_1, \hat{x}_3)$	0	1	2
	0	1	1	1
$\hat{x}_1$	1	1	0	0
	2	1	0	0

		$\hat{x}_3$		
	$\hat{b}_3(\hat{x}_1, \hat{x}_3)$	0	1	2
	0	2	2	1
$\hat{x}_1$	1	2	2	1
	2	0	0	0

(c)

Figura 8: (a) Exemplo de rede regulatória em notação de grafos, onde as arestas estão rotuladas com o número de limiar e o signo de sua influência. (b) Equações lógicas correspondentes ao grafo e (c) possíveis definições das funções lógicas.

somente uma das influências inibidoras for operativa, então o gene 2 é moderadamente expresso. Isto é representado pelo valor 1 para a imagem de $\hat{x}_2$. Em geral, varias funções lógicas serão consistentes com as restrições dos limiares. Qual função lógica é escolhida pode ser motivado por considerações biológicas ou pode ser uma suposição que reflete a incerteza sobre a estrutura do sistema que está sendo estudado.

As equações lógicas são o ponto de partida para o análise do sistema regulatório, em particular a determinação de seus *estados lógicos estacionários.* Um estado lógico estacionário ocorre quando o estado lógico do sistema iguala a sua imagem, isto é,

$$\hat{\mathbf{X}} = \hat{\mathbf{x}} \tag{4}$$

Como o número de estados lógicos é finito, devido à discretização realizada pela introdução de variáveis lógicas, é possível procurar exaustivamente os estados lógicos. No caso das equações lógicas da Figura 8, o estado lógico $[2 \quad 0 \quad 1]'$ é o único estado estacionário entre os 18 estados lógicos possíveis. Os demais são *estados lógicos transitórios.*

Se o sistema regulatório está em um estado lógico transitório, ele fará uma transição para outro estado lógico [52, 53]. Devido a que um valor lógico se moverá no sentido de sua imagem, os possíveis estados sucessores de um estado lógico podem ser deduzidos comparando o valor de uma variável lógica com o de sua imagem. Supondo que duas variáveis lógicas não mudarão seu valor simultaneamente, no máximo podem ser alcançados n estados a partir de um estado atual. Naturalmente, se um estado lógico for estacionário, não tem nenhum estado sucessor, dado que as variáveis lógicas são iguais a sua imagem. Os estados lógicos e as transições entre eles podem ser organizados em um *grafo de transição de estados.* Em analises mais avançadas

das transições de estados, podem considerar-se atrasos de tempo ocasionados pela transcrição, tradução, e o transporte [53, 56].

2.5 Equações Diferenciais Ordinárias não Lineares

As *equações diferenciais ordinárias* (ODEs: "ordinary differential equations") foram usadas extensamente para analisar sistemas de regulação gênica. O formalismo modela as concentrações de RNAs, proteínas, e outras moléculas por variáveis dependentes do tempo com valores no conjunto dos números reais não negativos. As interações regulatórias tomam a forma de relações funcionais e diferenciais entre as variáveis de concentração. Mais especificamente, a regulação gênica é modelada por equações –*"rate equations"*– que expressam a *taxa* de produção de um componente do sistema como uma função das concentrações de outros componentes. Esta equações tem a forma matemática:

$$\frac{dx_i}{dt} = f_i(\mathbf{x}), \quad 1 \leq i \leq n, \tag{5}$$

onde $\mathbf{x} = [x_1, \ldots, x_n]' \geq \mathbf{0}$ é o vetor de concentrações de proteínas, mRNAs ou pequenas moléculas, e $f_i : \mathbb{R}^n \to \mathbb{R}$ uma função geralmente não-linear. A taxa de síntese de i é dependente das concentrações $\mathbf{x}$, possivelmente incluindo x_i, e f_i pode ser extendida para incluir as concentrações $\mathbf{u} \geq \mathbf{0}$ de componentes de entrada, por exemplo, nutrientes fornecidos externamente. Atrasos de tempo discretos, devidos ao tempo requerido para completar a transcrição, tradução, e a difusão até o lugar onde atúa a proteína, também podem ser representados

$$\frac{dx_i}{dt} = f_i(x_1(t - \tau_{i1}), \ldots, x_n(t - \tau_{in})), \quad 1 \leq i \leq n, \tag{6}$$

onde $\tau_{i1}, \ldots, \tau_{in} > 0$ denotam atrasos de tempo discretos. Alternativamente, podem usar-se integrais para modelar atrasos de tempo distribuídos [57, 58, 59, 60].

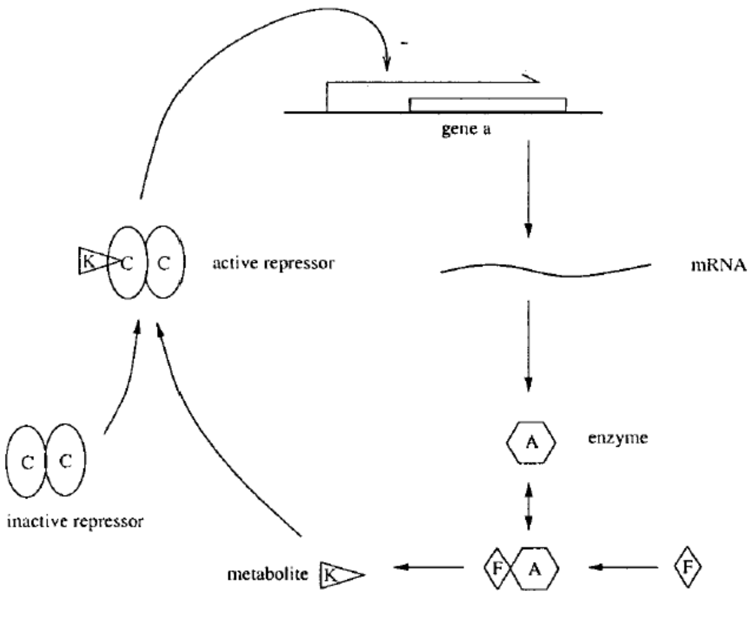
Foram desenvolvidos poderosos métodos matemáticos para modelar sistemas de reações bioquímicas por meio de *"rate equations"*, especialmente no contexto de processos metabólicos [61, 62]. Usando estes métodos podem construir-se modelos cinéticos de processos de regulação gênica especificando as funções f_i.

A Figura 9 mostra um modelo cinético para um processo simples de regulação gênica [63, 64]. O produto final de uma via metabólica co-inibe a expressão de um gene que codifica uma enzima que cataliza uma reação da via. Isto produz um laço de realimentação negativa que envolve a concentração de mRNA x_1, a concentração da enzima x_2 e a concentração do metabolito x_3.

Mais geralmente, são usadas equações da forma [65]:

$$\begin{aligned} \frac{dx_1}{dt} &= \kappa_{1n} r(x_n) - \gamma_1 x_1, \\ \frac{dx_i}{dt} &= \kappa_{i,i-1} x_{i-1} - \gamma_i x_i, \quad 1 < i \leq n. \end{aligned} \tag{7}$$

Os parâmetros $\kappa_{1n}, \kappa_{21} \ldots, \kappa_{n,n-1} > 0$ são constantes de produção e $\gamma_1, \ldots, \gamma_n$ são constantes de degradação. As equações expressam um equilíbrio entre o número de moléculas que aparecem e



(a)

$$\dot{x}_1 = \kappa_1 r(x_3) - \gamma_1 x_1$$
$$\dot{x}_2 = \kappa_2 x_1 - \gamma_2 x_2$$
$$\dot{x}_3 = \kappa_3 x_2 - \gamma_3 x_3$$

(b)

Figura 9: Exemplo de sistema regulatório genético envolvendo inibição produzida pelo produto final e (b) o seu modelo ODE. A é uma enzima e C uma proteína repressora, enquanto K e F são metabolitos; x_1, x_2 e x_3 representam as concentrações de mRNA a, proteína A, e metabolito K, respectivamente; κ_1, κ_2, κ_3 são constantes de produção, γ_1, γ_2, γ_3 constantes de degradação, e $r : \mathbb{R} \to \mathbb{R}$ uma função de regulação não-linear decrescente com valores de 0 a 1.

desaparecem por unidade de tempo. No caso de x_1, o termo de produção envolve uma *função de regulação* não-linear $r : \mathbb{R} \to \mathbb{R}$, enquanto as concentrações x_i, $1 < i \leq n$ aumentam linearmente com x_{i-1}. Para expressar que o produto metabólico é um co-repressor do gene, r deve ser uma função decrescente, ou seja, $\partial r / \partial x_n < 0$. Os termos $-\gamma_i x_i$, $1 \leq i \leq n$, indicam que as concentrações x_i decrescem, devido a degradação, difusão e diluição, com uma taxa proporcional às concentrações mesmas.

Uma função de regulação freqüentemente encontrada na literatura é a chamada *curva de Hill* (Figura 10(a)):

$$h^+(x_j, \theta_j, m) = \frac{x_j^m}{x_j^m + \theta_j^m} \tag{8}$$

com $\theta_j > 0$, o *limiar* para a influência regulatória de x_j sobre um gene alvo, e $m > 0$ um parâmetro de passo. A função varia de 0 a 1 e aumenta a medida que $x_j \to \infty$, de modo que um aumento em x_j tenderá a aumentar a taxa de expressão do gene (*ativação*). Para expressar que um aumento em x_j diminui a taxa de expressão (*inibição*), como em (8), a função de regulação $h^+(x_j, \theta_j, m)$ é substituída por $h^-(x_j, \theta_j, m) = 1 - h^+(x_j, \theta_j, m)$. Para $m > 1$, as curvas de Hill tem forma de sigmoide, o qual concorda com evidência experimental de expressão gênica. A

Figura 10 mostra outros dois exemplos de funções de regulação.

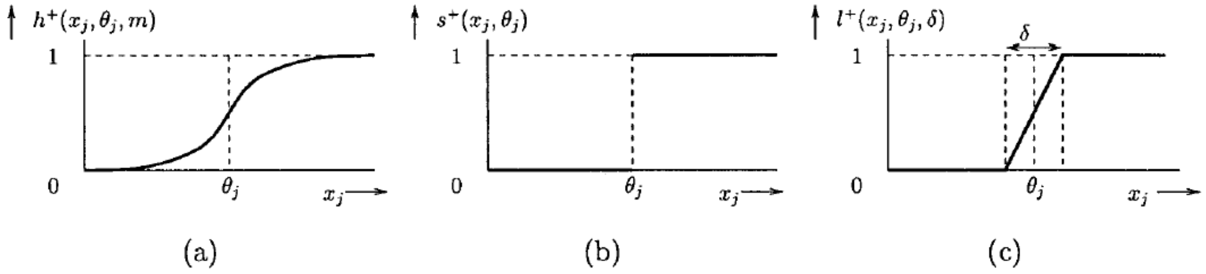


Figura 10: Exemplos de funções reguladoras: (a) função de Hill h^+, (b) função Heaviside ou degrau s^+, e (c) função logoide l^+

Devido à não-linearidade de f_i, a solução analítica das equações 5 normalmente não é possível. Em casos especiais, podem ser estabelecidas propriedades qualitativas das soluções, tais como o número e a estabilidade de estados estacionários e a ocorrência de ciclos limites. Isto é ilustrado por estudos que investigam a relação entre laços de realimentação e a dinâmica dos sistemas regulatórios [60, 65, 66, 67]. No caso de um laço de realimentação negativa, como na Figura 9, o sistema pode se aproximar ou oscilar em torno de um único estado estacionário. Na presença de um laço de realimentação positiva, pelo contrário, o sistema tende a se estabelecer em um de dois estados estáveis, dependendo de se o estado inicial está outro em um ou lado de uma separatriz.

De fato, um laço de realimentação negativa é uma condição necessária para periodicidade estável e um laço de realimentação positiva para multi-estacionaridade [68, 69]. A periodicidade estável dos laços de realimentação negativa pode ser interpretada como representando homeostase, enquanto a multi-estacionaridade ocasionada por um laço de realimentação positiva fornece um sugestivo paralelo com os processos de desenvolvimento observados.

Quando não for possível achar a solução analítica das equações, pode-se recorrer a técnicas numéricas. Na simulação numérica, a solução exata das equações é aproximada calculando valores aproximados $\mathbf{x}_0, \ldots, \mathbf{x}_m$ em instantes de tempo consecutivos $t_0, \ldots, t_m$ [70]. A simulação do funcionamento de uma rede regulatória é freqüentemente complementada com uma *análise de bifurcação* para investigar a sensibilidade de estados estacionários e ciclos limites aos valores dos parâmetros.

Um problema referente ao uso de técnicas numéricas é a falta de medições *in vivo* ou *in vitro* dos parâmetros cinéticos das equações. Valores numéricos dos parâmetros estão disponíveis somente para um punhado de sistemas bem estudados. Em modelos de ciclo celular, na maioria dos casos os valores dos parâmetros devem ser escolhidos de modo que os modelos consigam reproduzir o comportamento qualitativo observado.

A disponibilidade crescente de dados de expressão gênica pode aliviar algo o problema. A partir de medidas das variáveis de estado $\mathbf{x}$ em varias etapas de um processo de interesse, possivelmente sob uma variedade de condições experimentais, os valores dos parâmetros podem ser estimados com a ajuda de técnicas de identificação de sistemas [71].

2.6 Equações Diferenciais Lineares por Partes

Consideraremos agora um caso especial das equações (5) baseado em duas suposições simplificadoras. Primeiramente, os modelos simplificam os detalhes bioquímicos das interações regulatórias diretamente relacionando os níveis de expressão dos genes na rede. Segundo, o comportamento "ativo-inativo" dos genes, cuja expressão é regulada por curvas sigmoides contínuas, é aproximado por funções degrau descontínuas (Figura 10). As equações diferenciais são lineares por partes têm propriedades matemáticas favoráveis que facilitam a sua análise qualitativa. Os modelos de *Equações diferenciais lineares por partes* (PLDE: "Piecewise-linear differential equation") estão relacionados aos modelos de redes Booleanas e redes lógicas generalizadas, apesar de sutis diferenças causadas pela natureza discreta ao invés de contínua destes formalismos.

Consideraremos equações diferenciais lineares por partes da forma:

$$\frac{dx_i}{dt} = g_i(\mathbf{x}) - \gamma_i x_i, \quad 1 \leq i \leq n, \tag{9}$$

onde x_i denota a concentração celular do produto do gene i e $\gamma_i > 0$ a taxa de degradação de x_i [53, 72].

Na sua forma mais geral, a função $g_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ é definida como

$$g_i(\mathbf{x}) = \sum_{l \in L} \kappa_{il} b_{il}(\mathbf{x}) \geq 0, \tag{10}$$

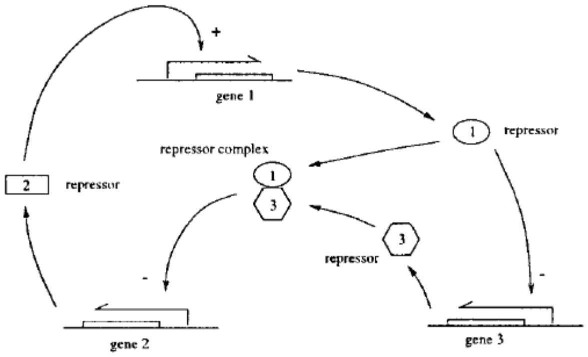
onde $\kappa_{il} > 0$ é um parâmetro de taxa de variação, $b_{il} : \mathbb{R}^n_{\geq 0} \to \{0,1\}$ uma função definida em termos de sumas e multiplicações de funções degrau, e L um conjunto possivelmente vazio de índices. A função b_{il} deve ser vista como o equivalente aritmético de uma função Booleana, que expressa as condições sob as que o gene é expresso a uma taxa κ_{il} [73, 74]. As condições são especificadas por meio de funções degrau s^+ e s^-, definidas por

$$s^+(x_j, \theta_j) = \begin{cases} 1, & x_j > \theta_j \\ 0, & x_j < \theta_j \end{cases}, \quad \text{e} \quad s^-(x_j, \theta_j) = 1 - s^+(x_j, \theta_j). \tag{11}$$

No exemplo da Figura 11, vemos que para o gene 1, $g_1(x_2) = \kappa_1 s^+(x_2, \theta_{21})$, o que significa que a proteína 1 é sintetizada a uma taxa κ_1, se a concentração da proteína 2 estiver abaixo da concentração limiar θ_{21}. O gene 2 é reprimido por um heterodímero composto pelas proteínas 1 e 3, de modo que ele é expresso a uma taxa κ_2, se a concentração de cada proteína fica abaixo do seu respectivo limiar ($g_2(x_1, x_3) = \kappa_2(1 - s^+(x_1, \theta_{11}) s^+(x_3, \theta_{31}))$). Podem usar-se combinações mas complexas de funções degrau para modelar efeitos combinados de varias proteínas regulatórias.

As equações 9 foram bem estudadas em biologia matemática e em outras áreas [53, 75, 76]. Considere um (hiper)cubo n-dimensional do espaço de fase definido da seguinte maneira:

$$0 \leq x_i \leq max_i > \max_{\mathbf{x} \geq \mathbf{0}} g_i(\mathbf{x})/\gamma_i, \quad 1 \leq i \leq n, \tag{12}$$



(a)

$$
\begin{aligned}
\dot{x}_1 &= \kappa_1\, s^+(x_2, \theta_{21}) - \gamma_1\, x_1 \\
\dot{x}_2 &= \kappa_2\, (1 - s^+(x_1, \theta_{11})\, s^+(x_3, \theta_{31})) - \gamma_2\, x_2 \\
\dot{x}_3 &= \kappa_3\, s^-(x_1, \theta_{12}) + \kappa_4\, s^-(x_3, \theta_{32}) - \gamma_3\, x_3
\end{aligned}
$$

(b)

Figura 11: (a) Exemplo de rede regulatória de três genes e (b) equações diferenciais não-lineares por partes; x_1, x_2, e x_3 representam concentrações de proteína o de mRNA, respectivamente, $\kappa_1, \ldots, \kappa_4$ constantes de produção, e $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{31}, \theta_{32}$ constantes de limiar.

onde supomos que para todas as concentrações limiar θ_{ik} da proteína codificada pelo gene i, vale que $\theta_{ik} < max_i$. Os (hiper)planos limiares $n-1$-dimensionais $x_i = \theta_{ik}$ dividem o n-cubo em ortantes. A Figura 12(a) mostra o cubo do espaço de fase e os ortantes correspondentes ao sistema da Figura 11.

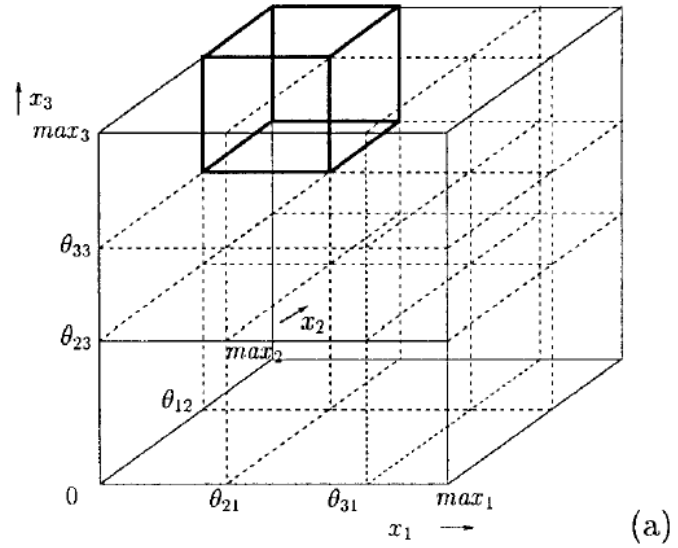
Em cada ortante do n-cubo, determinado pelo valor das funções degrau, (9) reduzen-se a equações diferenciais ordinárias com um termo de produção constante μ_i composto de parâmetros de taxa em b_i:

$$\dot{x}_i = \mu_i - \gamma_i x_i, \quad 1 \leq i \leq n. \tag{13}$$

A Figura 12(b) fornece um exemplo das equações de estado correspondentes ao ortante
$0 \leq x_1 < \theta_{21}, \quad \theta_{12} < x_2 \leq max_2, \quad$ e $\quad \theta_{33} < x_3 \leq max_3$.

Dado que as equações (13) são lineares e ortogonais é simples analisar o comportamento dentro de um ortante. Todas as trajetórias evoluem para um estado estacionário estável $\mathbf{x}^* = \mu/\lambda$, chamado *estado focal.* O estado focal de um ortante pode estar dentro ou fora do ortante. Se o estado focal fica fora do ortante, as trajetórias tendem a um ou vários dos planos limiares que limitam o ortante. As trajetórias podem continuar em um ortante adjacente, com um novo estado focal determinado pelas equações de estado para esse ortante. Se as trajetórias podem continuar num ortante adjacente dependerá da forma precisa na qual as equações (9) são definidas nos planos limiares, onde as funções degrau são discontínuas. Na literatura tem sido propostas varias alternativas [77, 78, 79], que vão desde restrições sobre a função g_i até soluções gerais baseadas em técnicas matemáticas sofisticadas.

As PLDEs (9) tem dois tipos de estados estacionários. Além dos estados estacionários *regulares* que ficam dentro dos ortantes, há estados estacionários *singulares* localizados sobre um ou mais dos planos limiares que separam os ortantes [75]. Os estados estacionários regulares são facilmente identificáveis, dado que eles são o estado focal de algum ortante no n-cubo. Os



$$\dot{x}_1 = \kappa_{12} - \gamma_1 x_1$$
$$\dot{x}_2 = -\gamma_2 x_2$$
$$\dot{x}_3 = \kappa_{31} - \gamma_3 x_3$$

(b)

Figura 12: (a) Cubo do espaço de fase para o modelo da Figura 11, dividido em $2 \cdot 3 \cdot 3 = 18$ ortantes pelos planos limiares.
(b) As equações de estado para o ortante $0 \leq x_1 < \theta_{21}$, $\quad \theta_{12} < x_2 \leq max_2$, e $\theta_{33} < x_3 \leq max_3$ (o ortante realçado en linhas grossas).

estados estacionários singulares, importantes devido a que são possíveis pontos homeostáticos do sistema regulatório, representam uma dificuldade para o formalismo linear por partes, devido às descontinuidades nos limiares. Algumas técnicas para identificar estes estados estacionários foram apresentadas em [75, 78, 79].

O comportamento global das PLDEs pode ser complexo e não é bem compreendido em geral. Continuações das trajetórias em vários ortantes podem originar oscilações para um estado estacionário singular situado na intersecção de planos limiares, ciclos ou ciclos limites [79, 80]. Estudos baseados em simulações numéricas mostraram que dinâmicas aperiódicas e caóticas podem ocorrer para $n \geq 4$ e são bastante comuns para dimensões mais elevadas e características particulares das funções g_i [81, 82].

Quanta informação se perde pela aproximação da regulação sigmoidal por funções degrau? Estudos feitos com simulações numéricas [83, 84] mostram que em a maioria de casos não há diferença nas propriedades qualitativas das soluções quando são usadas funções degrau em vez de funções de Hill.

Uma vantagem importante do uso de PLDEs é a simplicidade da análise matemática conferida pela forma das equações. Uma técnica similar proposta por [85] consiste em definir g_i em(9) por meio de *funções lei-de-potência*, originando modelos não-lineares da forma:

$$\dot{x}_i = \sum_{k=1}^{p} \kappa_{ik} \prod_{j=1}^{n} x_j^{r_{ijk}} - \sum_{k=1}^{p} \gamma_{ik} \prod_{j=1}^{n} x_j^{s_{ijk}}, \quad 1 \leq i \leq n, \tag{14}$$

onde os r_{ijk} (s_{ijk}) são ordem cinéticas dos processo elementares que contribuem à produção (degradação) dos x_i e κ_{ik} (γ_{ik}) são constantes de taxa para estes processos. Mudando para uma escala logarítmica, (14) se torna um sistema linear que é muito mais fácil de analisar que o sistema não-linear original [85]. Entre outras coisas, este formalismo foi aplicado ao estudo da eficácia funcional de diferentes tipos de acoplamentos em redes regulatórias [86].

2.7 Equações Estocásticas Principais ("Stochastic Master Equations")

Em princípio, as equações diferenciais permitem descrever a regulação gênica em grande detalhe, até o nível de passos de reação individuais como a ligação de um fator de transcrição a um sítio regulatório, ou a transcrição de um gene pelo avanço passo a passo da DNA polimerase ao longo da molécula de DNA (Figura 13). Entretanto, um número de suposições implícitas nos formalismos de equações diferenciais não são validas ao nível molecular.

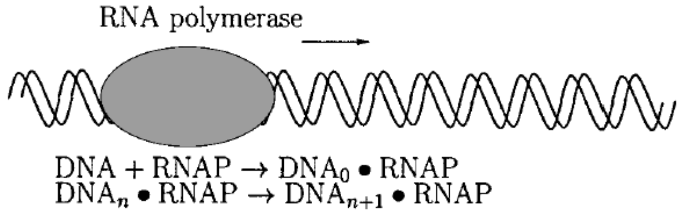


Figura 13: Exemplo de dois passos de reação envolvidos na transcrição de DNA

As equações diferenciais pressupõem que as concentrações de substâncias variam *continuamente* e *determinísticamente*, ambas as duas suposições podem ser questionáveis no caso de regulação gênica [87, 88, 89, 92]. Em primeiro lugar, o pequeno número de moléculas de alguns dos componentes do sistema regulatório compromete a suposição de continuidade. Pode haver somente umas poucas dezenas de moléculas de um fator de transcrição no núcleo da célula e uma única molécula de DNA que contem o gene. Segundo, a mudança determinística pressuposta pelo uso do operador diferencial d/dt pode ser questionável devido a flutuações no tempo em que ocorrem os eventos celulares, tal como o atraso entre começo e final da transcrição. Conseqüentemente, dois sistemas regulatórios com as mesmas condições iniciais podem finalmente estabelecer-se em estados diferentes, um fenômeno reforçado pelo pequeno número de moléculas envolvidas.

Em vez de fazer um exame contínuo e determinístico, alguns autores propuseram o uso de modelos de regulação gênica *discretos* e *estocásticos* [90, 91, 92]. As quantidades discretas de $\mathbf{X}$ de moléculas são tomadas como variáveis de estados e é introduzida uma distribuição de probabilidade conjunta $p(\mathbf{X}, t)$ para expressar a probabilidade de que no tempo t a célula contém X_1 moléculas da primeira espécie, X_2 moléculas da segunda espécie, etc. A evolução temporal da função $p(\mathbf{X}, t)$ pode ser especificada como:

$$p(\mathbf{X}, t + \Delta t) = p(\mathbf{X}, t) \left(1 - \sum_{j=1}^{m} \alpha_j \Delta t \right) + \sum_{j=1}^{m} \beta_j \Delta t, \tag{15}$$

onde m é o número de reações que podem ocorrer no sistema, $\alpha_j \Delta t$ a probabilidade de que a reação j ocorra no intervalo $[t, t+\Delta t]$ dado que o sistema está no estado $\mathbf{X}$ em t, e $\beta_j \Delta t$ a probabilidade de que a reação j trará ao sistema para o estado $\mathbf{X}$ desde outro estado em $[t, t+\Delta t]$ [92, 93]. Rearranjando (15), e tomando o limite quando $\Delta t \to 0$, obtemos a *equação principal* ("master equation") [94]:

$$\frac{\partial}{\partial t} p(\mathbf{X}, t) = \sum_{j=1}^{m} (\beta_j - \alpha_j p(\mathbf{X}, t)) \tag{16}$$

Comparemos esta equação com as equações de taxa (5) das equações diferenciais ordinárias não-lineares. Enquanto que (5) determinam como o estado do sistema muda ao longo do tempo, (16) descreve como a probabilidade de que o sistema está num determinado estado muda com o tempo. As variáveis de estado na formulação estocástica podem ser reformuladas como concentrações dividindo o número de moléculas X_i por um fator de volume.

Embora a equação principal forneça uma imagem intuitivamente clara dos processos estocásticos que governam a dinâmica de um sistema regulatório, é ainda mais difícil de resolver por meios analíticos do que as equações de taxa. Além disso, a simulação numérica do sistema é complicada pela forma de (16), que contem $n+1$ variáveis independentes: n variáveis discretas $\mathbf{X}$ e uma variável contínua t. Sob determinadas condições, a equação principal pode ser aproximada por equações diferenciais estocásticas, que consistem de uma equação diferencial ordinária do tipo (5) estendida com um termo de ruido [94, 95]. Não sempre é possível satisfazer as condições sob as quais esta aproximação é valida no caso de sistemas regulatórios genéticos.

Um enfoque alternativo seria não usar a equação principal e diretamente simular a evolução temporal do sistema regulatório. Este é o enfoque usado na *simulação estocástica* [92]. Basicamente, o algoritmo de simulação estocástica:

(i) determina *quando* ocorre a próxima reação e de *qual* tipo será, dado que o sistema está no estado $\mathbf{X}$ em t,

(ii) revisa o estado do sistema de acordo com esta reação, e

(iii) continua no estado seguinte.

São introduzidas as varíaveis estocásticas τ e ρ, as quais representam o intervalo de tempo até que a próxima reação ocorre e o tipo de reação, respectivamente. Em cada estado, é escolhido aleatoriamente um valor para τ e ρ de um conjunto de valores cuja função densidade de probabilidade $p(\tau, \rho)$ tem sido derivada dos mesmos princípios que aqueles subjacentes a equação principal (15). Isto garante que quando são realizadas um grande número de simulações estocásticas, a distribuição resultante para $\mathbf{X}$ em t se aproximará à distribuição implicada pela equação principal. Ou seja, enquanto que a equação principal trata das médias do comportamento, obtidas calculando as médias e as variâncias das X_is em t de $p(\mathbf{X}, t)$, a simulação estocástica fornece informação sobre comportamentos individuais.

A simulação estocástica resulta em aproximações mais próximas à realidade molecular da regulação gênica. Em primeiro lugar, o enfoque requer disponibilidade de conhecimento detalhado dos mecanismos da reação, incluindo estimativas da função densidade de probabilidade

$p(\tau, \rho)$. Além disso, a simulação estocástica é cara, devido ao grande número de simulações que necessitam ser realizadas para calcular um valor aproximado de $p(\mathbf{x}, t)$ e devido ao grande número de reações que necessitam ser analisadas em cada uma destas simulações. Se os custos são equilibrados pelos benefícios esperados depende do grau de granularidade à qual se deseja estudar os processos regulatórios. Numa escala de tempo maior, os efeitos estocásticos podem se nivelar, de modo que os modelos contínuos e determinísticos constituam uma boa aproximação [95].

2.8 Cadeias de Markov

Definições e Notação

Consideremos um alfabeto $S = \{s_1, s_2, \ldots, s_J\}$ e uma seqüência de variáveis aleatórias $X_0, X_1, \ldots, X_n, \ldots$, com valores em S. Os símbolos s_j no alfabeto são chamados *estados* e S é chamado o *espaço de estados*. Podemos codificar os estados dando ao estado s_j o rótulo j e tomar por conveniência $S = \{1, 2, \ldots, J\}$.

Uma seqüência de variáveis aleatórias $\{X_n\}_{n=0}^{\infty}$ é chamada uma *cadeia de Markov*, se para todo $n \geq 1$ e $j_0, \ldots, j_n \in S$,

$$P(X_n = j_n | X_0 = j_0, \ldots, X_{n-1} = j_{n-1}) = P(X_n = j_n | X_{n-1} = j_{n-1})). \tag{17}$$

A importância de uma cadeia de Markov está em que se $X_n = j_n$ é um evento futuro, então a probabilidade condicional deste evento dada a historia passada $X_0 = j_0, X_1 = j_1, \ldots, X_{n-1} = j_{n-1}$ depende somente do passado imediato $X_{n-1} = j_{n-1}$ e não do passado distante $X_0 = j_0, X_1 = j_1, \ldots, X_{n-2} = j_{n-2}$. A condição em (17) é conhecida como a *propriedade de Markov*.

Seja $\{X_n\}_{n=0}^{\infty}$ uma cadeia de Markov. Se $X_n = j_n$, dizemos que *a cadeia está no estado j no tempo n* ou que *a cadeia visita o estado j no tempo n*. As probabilidades condicionais

$$p_{i|j} = P(X_n = j | X_{n-1} = i), \qquad n \geq 1, \qquad i, j \in S \tag{18}$$

são supostas de ser independentes de n e são chamadas *probabilidades de transição de um passo (estacionárias)*. Uma cadeia de Markov com probabilidades de transição estacionárias é chamada de *homogênea*. Se a probabilidade condicional não é definida, fazemos $p_{i|j} = 0$.

Os números $p_{i|j}$ são por conveniência mostrados em forma de matriz

$$P = \left(p_{i|j}\right)_{i-1,j=1}^{J,J} \tag{19}$$

ou

$$P = \begin{pmatrix} p_{1|1} & p_{1|2} & \cdots & p_{1|J} \\ p_{2|1} & p_{2|2} & \cdots & p_{2|J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{J|1} & p_{J|2} & \cdots & p_{J|J} \end{pmatrix} \tag{20}$$

P é uma matriz $J \times J$ chamada *matriz de transição*. A linha i de P é a distribuição condicional de probabilidade de X_n dado que $X_{n-1} = i$. Portanto as seguintes probabilidades são verdadeiras:

$$p_{i|j} \geq 0, \qquad \sum_{j=1}^{J} p_{i|j} = 1. \tag{21}$$

Uma matriz com estas propriedades é chamada de *matriz estocástica*.

Cadeias de Markov de ordem k

Uma seqüência de variáveis aleatórias $\{X_n\}_{n=0}^{\infty}$ é chamada *cadeia de Markov de ordem k*, se para todo $n \geq 1$ e $j_0, j_1, \ldots, j_n \in S$,

$$\begin{aligned} &P(X_n = j_n | X_0 = j_0, X_1 = j_1, \ldots, X_{n-1} = j_{n-1}) \\ &= P(X_n = j_n | X_{n-k} = j_{n-k}, \ldots, X_{n-1} = j_{n-1}), \end{aligned} \tag{22}$$

para um inteiro positivo k.

Neste contexto a cadeia de Markov na definição básica (17) é chamada cadeia de Markov de primeira ordem. Cadeias de Markov de ordem maior que um são freqüentemente usadas na modelagem de genomas. Uma dificuldade com (22) é o número de parâmetros envolvidos.

Distribuição de Probabilidade Conjunta de uma Cadeia de Markov

Aqui atribuímos uma probabilidade à seqüência $j_0 \ldots j_n$ usando como modelo uma cadeia de Markov. Por sucessivas iterações da definição de probabilidade condicional e por sucessivos usos da propriedade de Markov (17)

$$\begin{aligned} &P(X_0 = j_0, \ldots, X_{n-1} = j_{n-1}, X_n = j_n) \\ &= P(X_n = j_n | X_0 = j_0, \ldots, X_{n-1} = j_{n-1}) P(X_0 = j_0, \ldots, X_{n-1} = j_{n-1}) \\ &= P(X_n = j_n | X_{n-1} = j_{n-1}) P(X_0 = j_0, \ldots, X_{n-1} = j_{n-1}) \\ &= p_{j_{n-1}|j_n} P(X_{n-1} = j_n | X_0 = j_0, \ldots, X_{n-2} = j_{n-2}) P(X_0 = j_0, \ldots, X_{n-2} = j_{n-2}) \\ &\vdots \\ &p_{j_{n-1}|j_n} \cdot p_{j_{n-2}|j_{n-1}} \cdots p_{j_0|j_1} \cdot p_{X_0}(j_0) \\ &p_{X_0}(j_0) \cdot p_{j_0|j_1} \cdots p_{j_{n-2}|j_{n-1}} \cdot p_{j_{n-1}|j_n}. \end{aligned} \tag{23}$$

Se algum dos $p_{j_{l-1}|j_l} = 0$ o resultado ainda é verdadeiro, dado que ambos os lados da equação são zero. Assim, isto prova que

Se $\{X_n\}_{n=0}^{\infty}$ é uma cadeia de Markov com probabilidade de transição estacionárias, então

$$P(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n) = \prod_{l=1}^{n} p_{j_{l-1}|j_l}. \tag{24}$$

Pode dizer-se que a propriedade anterior especifica completamente a família de distribuições de dimensão finita de $\{X_n\}_{n=0}^{\infty}$ de forma consistente e assim faz que $\{X_n\}_{n=0}^{\infty}$ seja um processo estocástico no sentido do teorema de Kolmogorov-Daniel [96], de modo que podemos escrever a equação anterior como

$$\{X_n\}_{n=0}^{\infty} \in \text{Markov}(P, p_{x_0}). \tag{25}$$

2.8.1 Equações de Chapman-Kolmogorov

Seja $\{X_n\}_{n=0}^{\infty} \in \text{Markov}(P, p_{x_0})$. Pode mostrar-se que as probabilidades condicionais

$$p_{i|j}(n) = P(X_{m+n} = j | X_m = i), \qquad n \geq 1, i, j \in S \tag{26}$$

são independentes. As probabilidades $p_{i|j}(n)$ são chamadas *probabilidades de transição de n passos.* Então

$$P(n) = \left(p_{i|j}(n)\right)_{i=1,j=1}^{J,J} \tag{27}$$

é a *matriz de transição de n passos.* Definimos

$$p_{i|j}(0) = \begin{cases} 1 & \text{se } j = 1 \\ 0 & \text{se } j \neq 1 \end{cases}$$

Equações de Chapman-Kolmogorov

Para todo $m, n \geq 1$ e $i, j \in S$ [17]:

$$p_{i|j}(m+n) = \sum_{k=1}^{J} p_{i|k}(m) \cdot p_{k|j}(n). \tag{28}$$

Em notação matricial esta equação pode ser escrita, com ajuda de (27), como o seguinte produto de matrizes

$$P(n+m) = P(m) \cdot P(n). \tag{29}$$

Mas temos [17]

$$P(n) = P^n. \tag{30}$$

Então a equação de Chapman-Kolmogorov (28) pode ser escrita como

$$P^{n+m} = P^m \cdot P^n. \tag{31}$$

2.8.2 Estacionaridade e Distribuições Invariantes

Seja a distribuição de X_0 denotada por $\pi(0)$, ou seja,

$$\pi(0) = (p_{X_0}(1), \ldots, p_{X_0}(J)). \tag{32}$$

que chamaremos de *distribuição inicial.* Denotemos

$$\pi(n) = (p(X_n = 1), \ldots, p(X_n = J)) \tag{33}$$

ao vetor $1 \times J$ de probabilidades de que a cadeia está no estado j no tempo n. Então usando propriedades básicas

$$p(X_n = j) = \sum_{k=1}^{J} p_{k|j} \cdot p(X_{n-1} = k).$$

Isto em notação matricial é

$$\pi(n) = \pi(n-1)P. \tag{34}$$

Uma cadeia de Markov $\{X_n\}_{n=0}^{\infty}$ é chamada de *estacionária* se a probabilidade $p(X_n = j)$ é independente de n para todo j no espaço de estados. Além disso, chamamos a uma distribuição π *uma distribuição invariante* com

$$\pi = (\pi_1, \dots, \pi_J),$$

se $p(X_0 = j) = \pi_j$ implica que $p(X_1 = j) = \pi_j$.

Seja $\{X_n\}_{n=0}^{\infty} \in \text{Markov}(P, \pi(0))$. Cada distribuição invariante satisfaz a equação [17]:

$$\pi = \pi P \tag{35}$$

com as restrições

$$\sum_{j=1}^{J} \pi_j = 1, \qquad \pi_j \geq 0.$$

Cada cadeia de Markov com um espaço finito tem *pelo menos uma* distribuição uniforme [17].

2.8.3 Cadeias de Markov Ergódicas

A cadeia de Markov $\{X_n\}_{n=0}^{\infty}$ é dita de ser *ergódica* se existe uma distribuição $a = (a_1, \dots, a_J)$ sobre o espaço de estados S tal que

$$\lim_{n \to \infty} \pi(n) = a \tag{36}$$

para qualquer distribuição inicial $\pi(0)$.

Evidentemente, a distribuição limite a deve ser única.

Seja $\{X_n\}_{n=0}^{\infty} \in \text{Markov}(P, p_{X_0})$. Suponhamos que

$$\lim_{n \to \infty} \pi(n) = a \tag{37}$$

onde $a = (a_1, \dots, a_J)$ é uma distribuição de probabilidade. Então a é uma distribuição invariante.

Prova: Tomando limites em (34) temos

$$\begin{aligned} a &= \lim_{n\to\infty} \pi(n) = \lim_{n\to\infty} \pi(n+1) \\ &= \lim_{n\to\infty} (\pi(n)P) = (\lim_{n\to\infty} \pi(n))P = aP. \end{aligned}$$

Aqui a troca de ordem entre multiplicação matricial e limite é permitida, dado que o espaço de estados é finito.

□

Portanto podemos dizer que uma cadeia de Markov ergódica se tornará assintoticamente estacionária.

Classificação do Espaço de Estados

(1) O estado $j \in S$ pode ser *alcançado* desde o estado $i \in S$, se existe um inteiro positivo n tal que

$$p_{i|j}(n) > 0.$$

(2) A matriz de transição P ou a cadeia $\{X_n\}_{n=0}^{\infty}$ é chamada *irredutível* se cada estado pode ser alcançado desde qualquer outro estado.

Pode-se provar que cada cadeia de Markov irredutível com probabilidades de transição estacionárias tem uma distribuição invariante *única* [17].

2.9 Aprendizado de Cadeias de Markov

Aplica-se aqui a modelagem Bayesiana [17] a uma seqüência (de treinamento) usando uma família de modelos de Markov. Um modelo de Markov é completamente especificado por uma matriz de transição e uma distribuição inicial. O aprendizado probabilístico necessita:

1) uma distribuição de probabilidade Markoviana a qual especifica a probabilidade de qualquer seqüência condicionada pela matriz de transição e a distribuição inicial;

2) uma "a priori" que expressa a incerteza a respeito da matriz de transição e a distribuição inicial.

Quando 1) é combinada de um modo conhecido com a seqüência de treinamento, obtemos a função de verossimilhança [17] da seqüência com respeito à família de modelos de Markov. A função de verossimilhança é combinada com 2) a traves da regra de Bayes para produzir uma *distribuição a posteriori* para os parâmetros da família de modelos de Markov. Usando 1) e 2) podemos calcular as distribuições (que permitem fazer predições) e comparar modelos por meio de fatores de Bayes. A comparação de familias de modelos está especialmente relacionada com encontrar a *ordem da cadeia de Markov*, técnica usada na modelagem de seqüências de DNA.

2.9.1 Máxima Verossimilhança de Cadeias de Markov

Seja $\underline{\theta}$ a matriz de probabilidades de transição

$$\underline{\theta} = \begin{pmatrix} \theta_{1|1} & \theta_{1|2} & \cdots & \theta_{1|J} \\ \theta_{2|1} & \theta_{2|2} & \cdots & \theta_{2|J} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{J|1} & \theta_{J|2} & \cdots & \theta_{J|J} \end{pmatrix} \tag{38}$$

Desejamos estimar o modelo $\underline{\theta}$ na família de modelos probabilísticos $p(\mathbf{x}|\underline{\theta})$ para uma seqüência (de treinamento) $\mathbf{x}$ de $n+1$ símbolos em S^{n+1}

$$\mathbf{x} = (j_0 j_1 \cdots j_n) \in S^{n+1}$$

Vimos que para cadeias de Markov

$$p(\mathbf{x}|\underline{\theta}) = P(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n|\theta) = \pi_{j_0}(0) \prod_{l=1}^{n} \theta_{j_{l-1}|j_l}. \tag{39}$$

Então propomos a

FAMÍLIA DE MODELOS:

CONDICIONADA EM $\pi_{j_0}(0)$ E $\Theta = \underline{\theta}$, OS SÍMBOLOS EM $\mathbf{x}$ SÃO A SAÍDA DE UMA CADEIA DE MARKOV $\{X_n\}_{n\geq 0}$, COM PROBABILIDADES DE TRANSIÇÃO ESTACIONÁRIAS $\underline{\theta}$.

□

Temos no máximo $J^2 - J$ parâmetros de transição e $J-1$ probabilidades iniciais para estimar usando os dados $\mathbf{x}$. Fazemos uma aproximação ao omitir a distribuição inicial $\pi(0)$ como parte do problema de estimação. Uma forma de pensar isto é que conhecemos (o fixamos a vontade) o símbolo inicial antecipadamente. Consideremos primeiro o estimador de máxima verossimilhança para todos los parâmetros de transição desconhecidos.

Como uma função de $\underline{\theta}$ para $\mathbf{x}$ fixado, a seguinte função de verossimilhança aproximada ou condicional é

$$L(\underline{\theta}) = \prod_{l=1}^{n} \theta_{j_{l-1}|j_l} \tag{40}$$

A função de verossimilhança logarítmica correspondente é

$$\mathcal{L}(\underline{\theta}) = \sum_{l=1}^{n} \theta_{j_{l-1}|j_l}. \tag{41}$$

Introduzimos uma notação para o número de vezes que vemos uma transição de i para j em $\mathbf{x} = (j_0 j_1 \ldots j_n)$:

$$n_{i|j} = \text{ o número de } l \text{ tal que } 1 \leq l \leq n, j_{l-1} = i, j_l = j.$$

Usando a contagem de freqüência $n_{i|j}$ podemos escrever a função de verossimilhança como

$$L(\underline{\theta}) = \prod_{i=1}^{J} \prod_{j=1}^{J} \theta_{i|j}^{n_{i|j}}. \tag{42}$$

Os $n_{i|j}$ são estatísticas suficientes para esta família de modelos. A verossimilhança logarítmica (41) é

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{J} \sum_{j=1}^{J} n_{i|j} \ln \theta_{i|j}. \tag{43}$$

Seja também

$$n_i = \text{ o número de } l \text{ tal que } 0 \leq l \leq n-1, j_l = i, \tag{44}$$

de modo que n_i é igual ao número de vezes que a seqüência $\mathbf{x}$ visita o estado i, excluindo a possível visita no tempo final. Então temos [17] que

O estimador de máxima verossimilhança $\hat{\theta}_{i|j}$ de $\theta_{i|j}$ é

$$\hat{\theta}_{i|j} = \frac{n_{i|j}}{n_i}, \tag{45}$$

para todo i e j.

2.9.2 Média de Modelos

Distribuições Posteriores para Linhas na Matriz de Transição

Suponhamos que nossa incerteza sobre as *linhas de* $\hat{\theta}$ em (38)

$$\theta_i = (\theta_{i|1}, \ldots, \theta_{i|J})$$

é modelada por *variáveis aleatórias independentes* que tem suas respectivas densidades de Dirichlet [17], formuladas com

$$Dir(\theta_i; \alpha_i \cdot q_{i|1}, \ldots, \alpha_i \cdot q_{i|J}) = \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma(\alpha_i q_{i|j})} \cdot \prod_{j=1}^{J} \theta_{i|j}^{\alpha_i q_{i|j} - 1}, \tag{46}$$

onde

$$\alpha_i > 0 \qquad q_{i|j} > 0 \qquad \sum_{j=1}^{J} q_{i|j} = 1.$$

Então usamos como densidades a priori simultâneas a **densidade multi-variada de Dirichlet** [17] ou

$$\prod_{i=1}^{J} Dir(\theta_i; \alpha_i \cdot q_{i|1}, \ldots, \alpha_i \cdot q_{i|J}) \tag{47}$$

Por conseguinte, a densidade a posteriori, em vista de (40), igual a

$$p(\underline{\theta}|\mathbf{x}) = \frac{\prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma(\alpha_i q_{i|j})} \cdot \prod_{j=1}^{J} \theta_{i|j}^{\alpha_i q_{i|j} - 1}}{p(\mathbf{x})} \tag{48}$$

onde $p(\mathbf{x})$ é a padronização que faz $p(\underline{\theta}|\mathbf{x})$ uma densidade de probabilidade em $\underline{\theta}$.

Probabilidade Preditiva

Queremos saber qual é a probabilidade

$$P(X_{n+1} = j|X_n = i; \mathbf{x}),$$

onde a notação indica que a probabilidade é baseada em uma dada seqüência de treinamento $\mathbf{x} = (j_0 j_1 \dots j_n) \in S_{n+1}$. Considerando a modelagem de Markov vista anteriormente, a resposta poderia ser

$$\hat{P}_{ML}(X_{n+1} = j|X_n = i; \mathbf{x}) = \hat{\theta}_{i|j}$$

inserindo o estimador de máxima verossimilhança (ML:"maximum likelihood") da probabilidade de transição.

Usando a seqüência $\mathbf{x}$ podemos tomar alguma densidade posterior para $\underline{\theta}$ e então prover uma nova matriz de transição fazendo a *média do modelo*

$$P^*(X_{n+1} = j|X_n = i; \mathbf{x}) = \int \theta_{i|j} p(\mathbf{x}|\underline{\theta}) d\underline{\theta} \tag{49}$$

usando (47)

$$P^*(X_{n+1} = j|X_n = i; \mathbf{x}) = \int \theta_{i|j} p(\mathbf{x}|\underline{\theta}) d\underline{\theta} = \frac{\prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma(\alpha_i q_{i|j})} I_{i|j}(r,s)}{p(\mathbf{x})} \tag{50}$$

onde

$$I_{i|j}(r,s) = \int \prod_{j=1}^{J} \theta_{r|s} \theta_{i|j}^{n_{i|J} q_{i|J} - 1} d\theta_i.$$

Aqui a integração é com respeito a todos os parâmetros em $\underline{\theta}$, que estão, no entanto, separado em seus respectivos domínios como linhas de $\underline{\theta}$. Usando formulas bem conhecidas para avaliar integrais de Dirichlet, obtemos a expressão

$$P^*(X_{n+1} = j|X_n = i; \mathbf{x}) = \frac{n_{s|r} + \alpha_s q_{s|r}}{n_s + \alpha_s}. \tag{51}$$

3 Referências

[1] Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Tutorial: Gene Expression Analysis and Genetic Network Modeling. Pacific Symposium on Biocomputing, 1999.

[2] Andy Wuensche. Genomic Regulation Modeled as a Network with Basins of Attraction Pacific Symposium on Biocomputing, 1998.

[3] Bas Dutilh. Analysis of data from microarray experiments, the state of the art in gene network reconstruction Literature thesis. Theoretical biology and Bioinformatics. Utrecht University. October 29, 1999

[4] Patrik D'haeseleer. Reconstructing Gene Networks from Large Scale Gene Expression Data. Ph.D. dissertation, University of New Mexico, 2000.

[5] Zoltan Szallasi. Genetic Network Analysis in Light of Massively Parallel Biological Data Acquisition Pacific Symposium on Biocomputing 4:5-16, 1999.

[6] S. Liang, S. Furhman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing, 3:18 29, 1998.

[7] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. Pacific Symposium on Biocomputing, 4:17 28, 1999.

[8] R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity, 1:45 63, 1996.

[9] R. Somogyi, S. Fuhrman, M. Askenazi, and A. Wuensche. The gene exression matrix: towards the extraction of genetic network architectures. Nonlinear Analysis, Theory, Methods and Applications, Vol. 30, No. 3, pp. 1815-1824, 1997.

[10] M. Wahde and J. Hertz. Coarse-grained reversed engineering of genetic regulatory networks. BioSystems 55, 129-136, Elsevier, 2000.

[11] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. Pacific Symposium on Biocomputing, 4:29 40, 1999.

[12] T.E. Ideker, V. Thorsson, and R.M. Karp. Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design. Pacific Symposium on Biocomputing 5:302-313, 2000.

[13] S. A. Kauffman. The Origins of Order. Self-Organization and Selection in Evolution. Oxford University Press, 1993.

[14] Ilya Shmulevich, Edward R. Dougherty, Seungchan Kim and Wei Zhang. Probabilistic Boolean Networks: a rule based uncertainty model for gene regulatory networks. Bioinformatics Vol. 18 no. 0 2002 Pages 1-14.

[15] Armelin, H. A., Barrera, J., Dougherty, E. R., Gubitoso, M. D., Ferreira, J. E., Hirata, N.S.T., Neves, E.J., A simulator for gene expression networks, In: Proc. SPIE Microarrays: Optical Technologies and Informatics, San Jose: SPIE, v.4266. p.248 ? 259, 2001.

[16] Barrera, J., Martin, P. A., Dougherty, E. R., Gubitoso, M. D., Identification of input-free finite lattice dynamical systems under envelope constraints, In: International Symposium on Mathematical Morphology, 2002, Sydney. ISMM-02, 2002.

[17] Timo Koski. Hidden Markov Models for Bioinformatics. Kluwer Academic Publishers, Dordrecht Hardbound, ISBN 1-4020-0135-5, November 2001.

[18] van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D., and Wodak, S.J. 2000. Representing and analysing molecular and cellular function using the computer. Biol. Chem. 381, 921 935.

[19] Karp, P.D., Riley, M., Paley, S.M., Pellegrini-Toole, A., and Krummenacker, M. 1999a. EcoCyc: Encyclopedia of Escherichia coli genes and metabolism. Nucl. Acids Res. 27(1), 55 58.

[20] Kolpakov, F.A., Ananko, E.A., Kolesov, G.B., and Kolchanov, N.A. 1999. GeneNet: A gene network database and its automated visualization. Bioinformatics 14(6), 529 537.

[21] Samsonova, M.G., Savostyanova, E.G., Serov, V.N., Spirov, A.V., and Reinitz, J. 1998. GeNet: A database of genetic networks. In Proc. First Int. Conf. Bioinformatics Genome Regul. Struct., BGRS 98, Novosibirsk, ICG.

[22] Jacq, B., Horn, F., Janody, F., Gompel, N., Serralbo, O., Mohr, E., Leroy, C., Bellon, B., Fasano, L., Laurenti, P., and Röder, L. 1997. GIF-DB: A WWW database on gene interactions involved in Drosophila melanogaster development. Nucl. Acids Res. 25(1), 67 71.

[23] Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucl. Acids Res. 28(1), 27 30.

[24] Euzenat, J., Chemla, C., and Jacq, B. 1997. A knowledge base for D. melanogaster gene interactions involved in pattern formation. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, eds. Proc. Fifth Int. Conf. Intell. Syst. Mol. Biol. (ISMB-97), 108 119, Menlo Park, CA. AAAI Press.

[25] Salgado, H., Santos, A., Garza-Ramos, U., van Helden, J., Díaz, E., and Collado-Vides, J. 2000. RegulonDB (version 2.0): A database on transcriptional regulation in Escherichia coli. Nucl. Acids Res. 27(1), 59 60.

[26] Samsonova, M., and Serov, V.N. 1999. NetWork: An interactive interface to the tools for analysis of genetic network structure and dynamics. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB 99), vol. 4, 102 111, Singapore, World Scienti c Publishing.

[27] Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., Rechenmann, F., and Jacq, B. 1999. Grasping at molecular interactions and genetic networks in Drosophila melanogaster. Nucleic Acids Res. 27(1), 89 94. Sánchez, L., and Thieffry, D. 2001. A logical analysis of the Drosophila gap genes. J. Theor. Biol. 211, 115 141.

[28] Arkin, A., and Ross, J. 1995. Statistical construction of chemical reaction mechanisms from measured time-series. J. Phys. Chem. 99, 970 979.

[29] Arkin, A., Shen, P., and Ross, J. 1997. A test case of correlation metric construction of a reaction pathway from measurements. Science 277, 1275 1279.

[30] Chen, T., Filkov, V., and Skiena, S.S. 1999. Identifying gene regulatory networks from experimental data. In Proc. 3rd Ann. Int. Conf. Comp. Mol. Biol. (RECOMB 99), 94 103, New York, ACM Press.

[31] Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95, 717 728.

[32] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S.H. 2000. Functional discovery via a compendium of expression pro les. Cell 102(1), 109 126.

[33] Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M., and Shapiro, L. 2000. Global analysis of the genetic network controlling a bacterial cell cycle. Science 290, 2144 2148.

[34] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. 1998. Comprehensive identi cation of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9, 3273 3297.

[35] Zien, A., Küffner, R., Zimmer, R., and Lengauer, T. 2000. Analysis of gene expression data with pathway scores. In R. Altman, T.L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck, and H. Weissig, eds. Proc. 8th Int. Conf. Intell. Syst. Mol. Biol. (ISMB 2000), 407 417, Menlo Park, CA, AAAI Press.

[36] Friedman, N., Linial, M., Nachman, I., and Pe er, D. 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601 620.

[37] Heckerman, D. 1998. A tutorial on learning with Bayesian networks. In M.I. Jordan, ed. Learning in Graphical Models, Kluwer, Dordrecht.

[38] Pe er, D., Regev, A., Elidan, G., and Friedman, N. 2001. Inferring subnetworks from perturbed expression pro les. Bioinformatics 17(S1), S215 S224.

[39] Murphy, K., and Mian, S. 1999. Modelling gene expression data using dynamic Bayesian networks. Technical report, University of California at Berkeley, Berkeley, CA. Unpublished report.

[40] Platzer, U., Gumbel, M., and Meinzer, H.-P. 2001. Simulation of genetic networks controlling cell differentiation. Poster at ISMB Satellite Meeting on Computer Modeling of Cellular Processes, SIGSIM01.

[41] Jackson, E.R., Johnson, D., and Nash, W.G. 1986. Gene networks in development. J. Theor. Biol. 119(4), 379 396.

[42] Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S. 1998a. Identi cation of gene regulatory networks by strategic gene disruptions and gene overexpressions. In H. Karloff, ed., Proc. 9th Ann. ACM-SIAM Symp. Discr. Algorithms (SODA-98), 695 702, Society for Industrial and Applied Mathematics, Philadelphia.

[43] Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S. 1998b. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. In Genome Informatics, vol. 9, 151 160.

[44] Karp, R.M., Stoughton, R., and Yeung, K.Y. 1999b. Algorithms for choosing differential gene expression experiments. In Proc. 3rd Ann. Int. Conf. Computational Mol. Biol. (RECOMB 99), 208 217, New York, ACM Press.

[45] Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., and Eguch, Y. 2001. Development of a system for the inference of large scale genetic networks. In R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB 01), vol. 6, 446 458, Singapore, World Scienti c Publishing.

[46] Noda, K., Shinohara, A., Takeda, M., Matsumoto, S., Miyano, S., and Kuhara, S. 1988. Finding genetic network from experiments by weighted network model. In Genome Informatics, vol. 9, 141 150.

[47] Kauffman, S.A. 1974. The large-scale structure and dynamics of gene control circuits: An ensemble approach. J. Theor. Biol. 44, 167 190.

[48] Kauffman, S.A. 1991. Antichaos and adaptation. Sci. Am. 265, 78 84.

[49] Kauffman, S.A. 1993. The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press, New York.

[50] Szallasi, Z., and Liang, S. 1998. Modeling the normal and neoplastic cell cycle with realistic Boolean genetic networks : Their application for understanding carcinogenesis and assessing therapeutic strategies. In R.B. Altman, A.K. Dunker, L. Hunter, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB 98), vol. 3, 66 76, Singapore, World Scientific Publishing.

[51] Weisbuch, G. 1986. Networks of automata and biological organization. J. Theor. Biol. 121, 255 267.

[52] Thomas, R. 1991. Regulatory networks seen as asynchronous automata: A logical description. J. Theor. Biol. 153, 1 23.

[53] Thomas, R., and d Ari, R. 1990. Biological Feedback, CRC Press, Boca Raton, FL.

[54] Thomas, R., Thieffry, D., and Kaufman, M. 1995. Dynamical behaviour of biological regulatory networks: I. Biological rool of feedback loops and practical use of the concept of the loop-characteristic state. Bull. Math. Biol. 57(2), 247 276.

[55] van Ham, P. 1979. How to deal with more than two levels. In R. Thomas, ed. Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems, vol. 29 of Lecture Notes in Biomathematics, 326 343, Springer-Verlag, Berlin.

[56] Thomas, R. 1983. Logical vs. continuous description of systems comprising feedback loops: The relation between time delays and parameters. Stud. Phys. Theor. Chem. 28, 307 321.

[57] Landahl, H.D. 1969. Some conditions for sustained oscillations in biochemical chaines. B. Math. Biophys. 31, 775 787.

[58] Mahaffy, J.M. 1984. Cellular control models with linked positive and negative feedback and delays: I. The models. J. Math. Biol. 106, 89 102.

[59] MacDonald, N. 1989. Biological Delay Systems: Linear Stability Theory, Cambridge University Press, Cambridge.

[60] Smolen, P., Baxter, D.A., and Byrne, J.H. 2000. Modeling transcriptional control in gene networks: Methods, recent results, and future directions. Bull. Math. Biol. 62, 247 292.

[61] Heinrich, R., and Schuster, S. 1996. The Regulation of Cellular Systems, Chapman and Hall, New York.

[62] Voit, E.O. 2000. Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists, Cambridge University Press, Cambridge.

[63] Goodwin, B.C. 1963. Temporal Organization in Cells, Academic Press, New York.

[64] Goodwin, B.C. 1965. Oscillatory behavior in enzymatic control processes. In G. Weber, ed. Advances in Enzyme Regulation, 425 438. Pergamon Press, Oxford.

[65] Tyson, J.J., and Othmer, H.G. 1978. The dynamics of feedback control circuits in biochemical pathways. Prog. Theor. Biol. 5, 1 62.

[66] Cherry, J.L., and Adler, F.R. 2000. How to make a biological switch. J. Theor. Biol. 203, 117 133.

[67] Wolf, D.M., and Eeckman, F.H. 1998. On the relationship between genomic regulatory element organization and gene regulatory dynamics. J. Theor. Biol. 195, 167 186.

[68] Gouzé, J.-L. 1998. Positive and negative circuits in dynamical systems. J. Biol. Syst. 6(1), 11 15.

[69] Thomas, R., and Kaufman, M. 2001. Multistationarity, the basis of cell differentiation and memory: I. Structural conditions of multistationarity and other nontrivial behavior. Chaos 11(1), 170 179.

[70] Lambert, J.D. 1991. Numerical Methods for Ordinary Differential Equations. Wiley, Chichester.

[71] Ljung, L. 1999. System Identi cation: Theory for the User, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

[72] Mestl, T., Plahte, E., and Omholt, S.W. 1995a. A mathematical framework for describing and analysing gene regulatory networks. J. Theor. Biol. 176, 291 300.

[73] Plahte, E., Mestl, T., and Omholt, S.W. 1998. A methodological basis for description and analysis of systems with complex switch-like interactions. J. Math. Biol. 36, 321 348.

[74] Snoussi, E.H. 1989. Qualitative dynamics of piecewise-linear differential equations: A discrete mapping approach. Dynam. Stabil. Syst. 4(3 4), 189 207.

[75] Snoussi, E.H., and Thomas, R. 1993. Logical identi cation of all steady states: The concept of feedback loop characteristic states. Bull. Math. Biol. 55(5), 973 991.

[76] Edwards, R., Siegelmann, H.T., Aziza, K., and Glass, L. 2001. Symbolic dynamics and computation in model gene networks. Chaos 11(1), 160 169.

[77] Glass, L. 1977. Global analysis of nonlinear chemical kinetics. In B. Berne, ed. Statistical Mechanics, Part B: Time Dependent Processes, 311 349, Plenum Press, New York.

[78] Gouzé, J.-L., and Sari, T. 2001. A class of piecewise linear differential equations arising in biological models. Technical Report RR-4207, INRIA Sophia-Antipolis, Sophia-Antipolis.

[79] Plahte, E., Mestl, T., and Omholt, S.W. 1994. Global analysis of steady points for systems of differential equations with sigmoid interactions. Dynam. Stabil. Syst. 9(4), 275 291.

[80] Mestl, T., Plahte, E., and Omholt, S.W. 1995b. Periodic solutions in systems of piecewise-linear differential equations. Dynam. Stabil. Syst. 10(2), 179 193.

[81] Mestl, T., Bagley, R.J., and Glass, L. 1997. Common chaos in arbitrarily complex feedback networks. Phys. Rev. Lett. 79(4), 653 656.

[82] Lewis, J.E., and Glass, L. 1992. Nonlinear dynamics and symbolic dynamics of neural networks. Neural Comput. 4, 621 642.

[83] Glass, L., and Kauffman, S.A. 1973. The logical analysis of continuous non-linear biochemical control networks. J. Theor. Biol. 39, 103 129.

[84] Glass, L., and Pérez, R. 1974. Limit cycle oscillations in compartmental chemical systems. J. Chem. Phys. 61(12), 5242 5249.

[85] Savageau, M.A. 1996b. Power-law formalism: A canonical nonlinear approach to modeling and analysis. In V. Lakshmikantham, ed. Proc. 1st World Congr. Nonlinear Anal., vol. 4, 3323 3334, Berlin, De Gruyter.

[86] Savageau, M.A. 1996a. A kinetic formalism for integrative molecular biology: Manifestation in biochemical systems theory and use in elucidating design principles for gene circuits. In J. Collado-Vides, B. Magasanik, and T.F. Smith, eds. Integrative Approaches to Molecular Biology, 115 146, MIT Press, Cambridge, MA.

[87] Gibson, M.A., and Mjolsness, E. 2001. Modeling the activity of single genes. In J.M. Bower and H. Bolouri, eds. Computational Modeling of Genetic and Biochemical Networks, 1 48. MIT Press, Cambridge, MA.

[88] McAdams, H.H., and Arkin, A. 1999. It s a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15(2), 65 69.

[89] Szallasi, Z. 1999. Genetic network analysis in light of massively parallel biological data acquisition. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, eds. Proc. Pac. Symp. Biocomput. (PSB 99), vol. 4, 5 16, Singapore, World Scientific Publishing.

[90] Arkin, A., Ross, J., and McAdams, H.A. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage ¸-infected Escherichia coli cells. Genetics 149, 1633 1648.

[91] Morton-Firth, C.J., and Bray, D. 1998. Predicting temporal uctuations in an intracellular signalling pathway. J. Theor. Biol. 192, 117 128.

[92] Gillespie, D.T. 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81(25), 2340 2361.

[93] Gillespie, D.T. 1992. A rigorous derivation of the chemical master equation. Physica D 188, 404 425.

[94] van Kampen, N.G. 1997. Stochastic Processes in Physics and Chemistry, rev. ed., Elsevier, Amsterdam.

[95] Gillespie, D.T. 2000. The chemical Langevin equation. J. Chem. Phys. 113(1), 297 306.

[96] Gray, R.M. 1988. Probability, Random Processes and Ergodic Properties,. Springer Verlag, New York, Berlin, etc..