MAC 5701
Tópicos em Ciência da Computação
Plano de Estudos

Aluno: Nestor Walter Trepode
Orientador: Junior Barrera

1 Introdução

O presente plano de estudos a ser desenvolvido no primeiro semestre de 2003, tem como o objetivo a realização de uma pesquisa bibliográfica e estudo das redes de expressão gênica e dos diferentes modelos existentes para definir um modelo computacional adequado e focalizar o trabalho da tese de doutorado.

1.1 Projeto de Pesquisa de Doutorado

O projeto de pesquisa de doutorado do aluno tem como título "Identificação de Sistemas Dinâmicos de Reticulados" e refere-se à simulação e identificação de Sistemas Dinâmicos Finitos ou FDS ("Finite Dynamical Systems"), que constituem um modelo adequado para a representação de redes de expressão gênica. Mediante o uso de técnicas de otimização estatística será efetuada a identificação, a partir de amostras de dinâmicas do sistema ideal. Simulando o sistema identificado e comparando o seu comportamento com amostras do sistema ideal são avaliados os resultados obtidos. A identificação de sistemas divide-se em duas etapas: i-reconhecimento da arquitetura; ii-aprendizado da função de transição de estados. Ambos os aspectos serão considerados nos nossos estudos. Para lidar com amostras pequenas se efetuará a identificação com restrições da dinâmica de FDSs, estudando e desenvolvendo os aspectos teóricos e algoritmos necessários. Posteriormente, se realizará a identificação de sistemas estocásticos que representam mais adequadamente aos sistemas reais.

2 Área de Estudo

2.1 Redes de Expressão Gênica

Os cromossomos do núcleo da célula são estruturas replicáveis que contem o *DNA –ácido desoxiribonucléico–*, 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.

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.

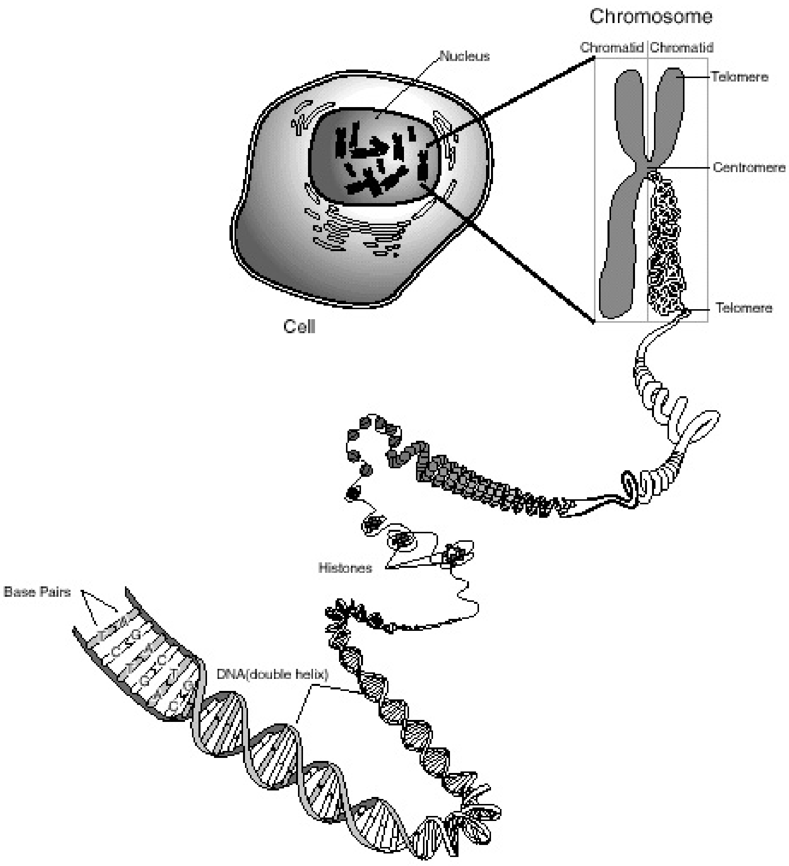

Figura 1: Cromossomo *(fonte: National Human Genome Institute).*

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.

2.2 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.

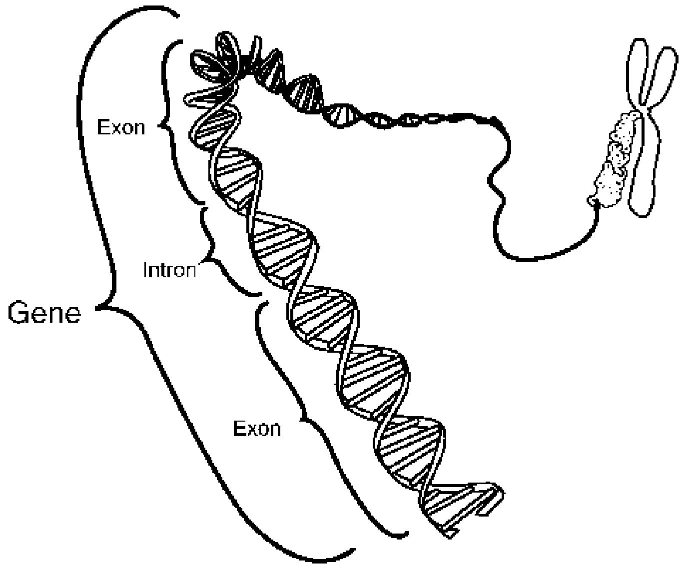

Figura 2: Gene *(fonte: National Human Genome Institute).*

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

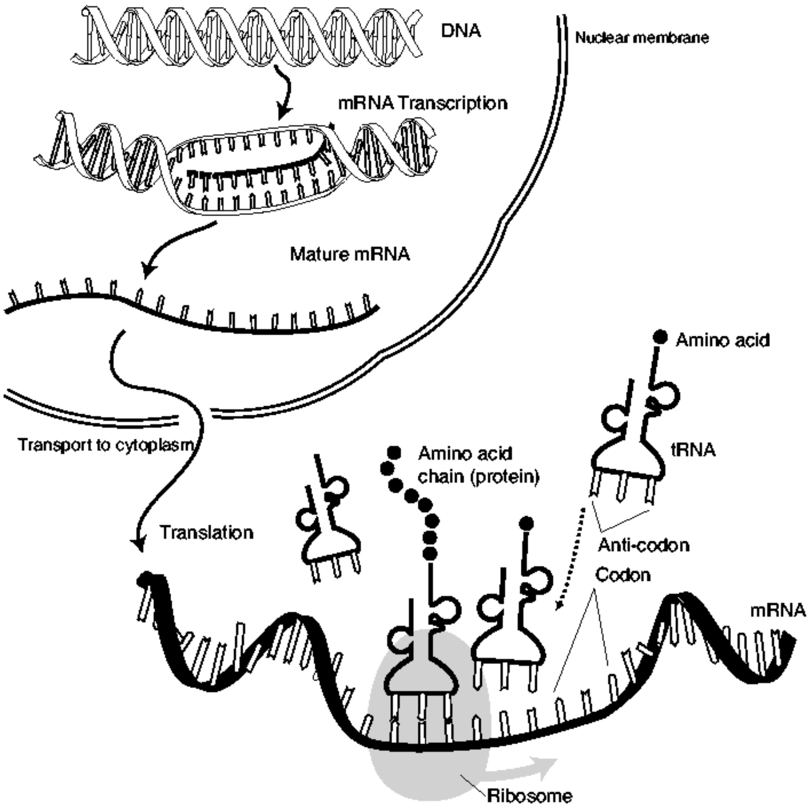

Figura 3: Expressão gênica *(fonte: National Human Genome Institute).*

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 insficiencia 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.

2.3 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

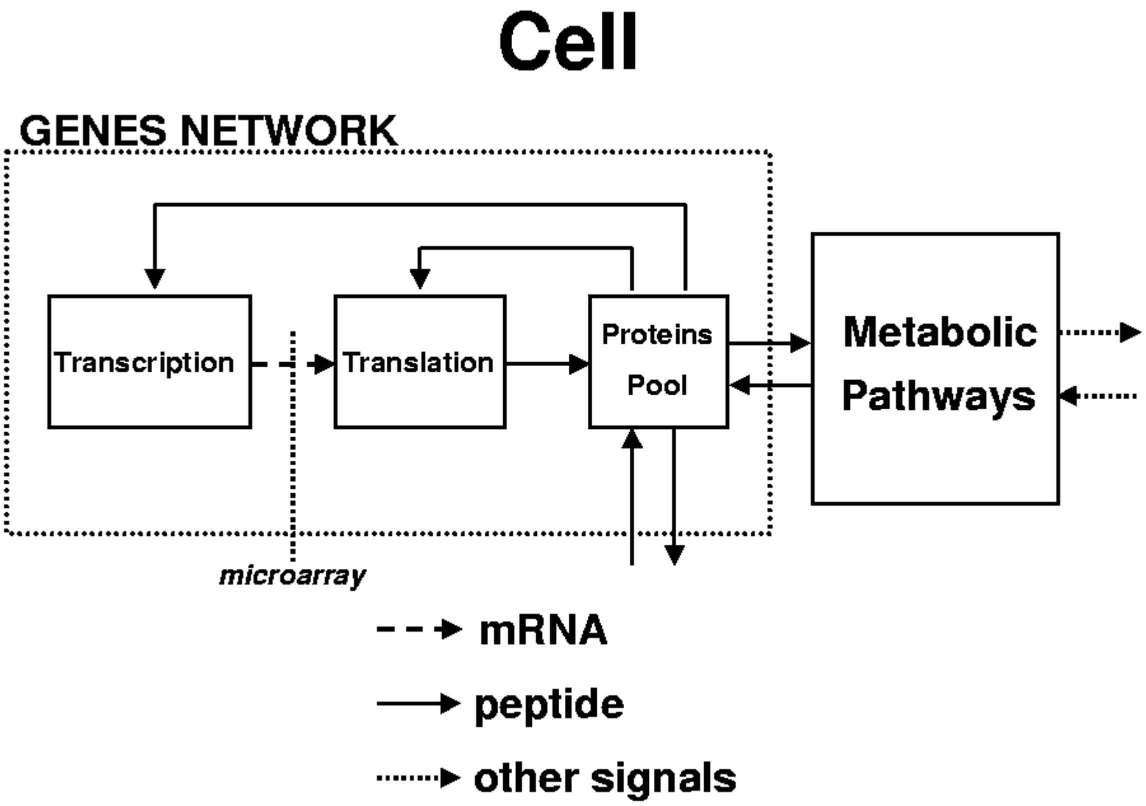

Figura 4: Rede de expressão gênica.

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 ambiguidades 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.

2.4 Identificação da Rede

Em *identificação de redes* o objetivo é construir um modelo da rede de interações regulatorias 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 regulatoria subjacente desconhecida. Usam geralmente um modelo paramétrico, e ajustam os parâmetros aos dados reais.

Se a arquitetura da rede regulatoria 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 premisas 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 trajectories 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.

3 Plano de Estudos

3.1 Tópicos

Continuaremos o nosso estudo do material obtido até agora. Será feita uma pesquisa bibliográfica e estudo referente a *Modelagem de Redes de Expressão Gênica* com o objetivo de definir e validar o modelo matemático computacional a ser utilizado no trabalho de pesquisa de doutorado.

Estudaremos as *cadeias de Markov* [17] que poderão ser utilizados na modelagem de redes de expressão gênica.

Parte do material estudado será apresentado em seminários.

3.2 Cronograma

- Março 2003: Busca de referências, planejamento do estudo, pesquisa.
- Abril 2003: Continuação da busca, estudo e resumo dos primeiros trabalhos.
- Maio 2003: Estudo e resumo dos demais trabalhos e da pesquisa realizada.
- Junho 2003: Elaboração do relatório final da disciplina.

4 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.