

Universidade Federal do Rio Grande
Programa de Pós-Graduação em Modelagem Computacional

Uma proposta para melhorar a convergência de MCMC

Nilzair Barreto Agostinho

Rio Grande, 9 de Junho de 2014

Universidade Federal do Rio Grande
Programa de Pós-Graduação em Modelagem Computacional

Uma proposta para melhorar a convergência de MCMC

Nilzair Barreto Agostinho

Banca Examinadora:

Prof. Dr. Adriano Velasque Werhli

C3 – FURG (Orientador)

Prof.^a Dr.^a Karina Machado

C3 – FURG

Prof. Dr. Leonardo Emmendorfer

C3 – FURG

Prof. Dr. Eder Simão

UNIFRA

*Dedico este trabalho aos meus pais Neir e Nilzabete que são
meu esteio, meu porto seguro, que sempre me dedicaram apoio
e amor incondicional.*

Agradecimentos

Primeiramente sou imensamente grata ao Senhor que tem derramado muitas bênçãos em minha vida e me ajuda em todos os momentos. Também gostaria de agradecer ao meu marido Edenilton que sempre me dá apoio e amor incondicional, aos meus pais Neir e Nilzabete, meu irmão Neir Jr e minha vó, aos meus amigos (Dania, Sylvia, Victor, Roberta, Rossana, Juliana, Dino, Mariane) que compreenderam quando muitas vezes me fiz ausente e não deixaram de me dedicar carinho e atenção, aos amigos que estiveram me dando suporte (Rejanete, Graziele, Jorge Cardoso, Gisele, Karina) durante períodos conturbados. E também gostaria de agradecer ao meu orientador Prof. Adriano Werhli por todo tempo que dedicou a mim ao me orientar neste trabalho e também por toda atenção, conselhos e amizade.

Conteúdo

Lista de Abreviaturas	v
Resumo	vii
1 Introdução	1
2 Metodologia	3
2.1 Redes Bayesianas (BN)	3
2.1.1 Aprendizagem de Redes Bayesianas (BNs)	5
2.2 Redes Bayesianas com <i>Markov Chain Monte Carlo</i> (MCMC) (Método Re- des Bayesianas- <i>Markov Chain Monte Carlo</i> (BNMCMC))	7
2.3 Combinando Conhecimento a Priori com a Expressão Gênica	8
2.4 Graphical Gaussian Models (<i>Graphical Gaussian Model</i> (GGM))	12
2.5 Informação Mútua	14
2.6 BNs Guiadas por GGMs	15
2.7 BNs Guiadas por <i>Mutual Information</i> - Informação Mútua (MI)	17
3 Simulações	18
3.1 Dados	18
3.1.1 Dados GeneNetWeaver - Ecoli	18
3.1.2 Dados Gaussianos de distribuição Gaussiana Multivariável	19
3.1.3 Dados Reais	19
3.2 Critérios de avaliação	20
3.3 Critérios de convergência	21
3.3.1 Avaliação visual Heurística	22

3.3.2	O método proposto de avaliação de convergência	22
3.3.3	Descrição Simulações	24
4	Resultados e Discussão	26
4.1	Método BNGGM	26
4.2	Método Modelo Bayesiano Hierárquico que utiliza MI (BNMI)	28
5	Conclusões e Discussão	40
A	Gráficos Curvas ROC	45
	Bibliografia	55

Lista de Figuras

2.1	Exemplo de rede Bayesiana	4
2.2	Incerteza Probabilística	6
2.3	Representação do MCMC	9
2.4	Modelo BNMCMC	9
2.5	Modelo Gráfico Probabilístico	10
2.6	Modelo Bayesiano Hierárquico (Modelo Bayesiano Hierárquico (BNGGM))	16
2.7	Modelo BNGGM / BNMI	17
3.1	Rede - <i>Gene Net Weaver</i> (GNW)	19
3.2	Rede - experimento citométrico	20
3.3	Método visual	23
3.4	Gráfico da Convergência	25
4.1	Média das áreas BNMCMC/BNGGM	27
4.2	Areas Ecoli BNGGM	29
4.3	Areas Gauss BNGGM	30
4.4	Areas Reais BNGGM	31
4.5	Convergências BNMCMC/BNGGM	32
4.6	Média das áreas BNMCMC/BNMI	35
4.7	Areas Ecoli BNMI	36
4.8	Areas Gauss BNMI	37
4.9	Areas Reais BNMI	38
4.10	Convergências BNMCMC/BNGGM	39
A.1	AUCs Ecoli/BN-MCMC	46

A.2	AUCs Gauss/BN-MCMC	47
A.3	AUCs Real/BN-MCMC	48
A.4	AUCs Ecoli/BNGGM	49
A.5	AUCs Gauss/BNGGM	50
A.6	AUCs Real/BNGGM	51
A.7	AUCs Ecoli/BNMI	52
A.8	AUCs Gauss/BNMI	53
A.9	AUCs Real/BNMI	54

Lista de abreviaturas

BDe *Discrete Multinomial Bayesian scoring metric*

BGe *Continuous Bayesian Gaussian scoring metric*

BN Rede Bayesiana

BNs Redes Bayesianas

BNGGM Modelo Bayesiano Hierárquico

BNMCMC Método Redes Bayesianas-*Markov Chain Monte Carlo*

BNMI Modelo Bayesiano Hierárquico que utiliza MI

BoN Redes Booleanas

ROC Curva de Características do Operador Receptor

DAG Grafo Acíclico Direcionado

DBN Rede Bayesina Dinâmica

DBNs Redes Bayesinas Dinâmicas

ED Equações Diferenciais

EDA Equações Diferenciais Acopladas

GGM *Graphical Gaussian Model*

GGMs *Graphical Gaussian Models*

GNW *Gene Net Weaver*

GRN Redes Regulatórias Genéticas

MCMC *Markov Chain Monte Carlo*

MI *Mutual Information* - Informação Mútua

PBN Redes Booleanas Probabilísticas

TP Verdadeira Positiva

FP Falsa Positiva

TN Verdadeira Negativa

FN Falsa Negativa

AUC Área abaixo da Curva de Características do Operador Receptor (ROC)

AUCs Áreas abaixo da ROC

Resumo

Atualmente na área de Sistemas Biológicos vêm-se trabalhando para o desenvolvimento de ferramentas que possam auxiliar a obter um maior conhecimento sobre as interações moleculares em um organismo. Pesquisas tem sido desenvolvidas tendo como objetivo principal obter um detalhamento da organização funcional de Sistemas Biológicos. As redes regulatórias genéticas tem sido utilizadas como um ferramental para dispor mapas detalhados sobre as interações entre os genes e consequentemente sobre as interações moleculares e a organização funcional. Porém, estas redes regulatórias são altamente complexas e o processo para inferi-las é custoso computacionalmente.

Neste trabalho optou-se por trabalhar com redes Bayesianas devido à sua natureza probabilística e flexibilidade. As redes Bayesianas foram amostradas através do MCMC que garante que haja uma convergência para distribuição posterior. Porém, na prática o MCMC é relativamente lento e sem garantia de alcançar a convergência. Desta forma propõe-se utilizar a saída de métodos mais rápidos e menos precisos como GGM como entrada para o método MCMC. Sendo assim, o presente trabalho propõe o desenvolvimento de um BNGGM permitindo que a amostragem MCMC seja guiada por um método mais rápido. Sendo assim, buscou-se então contribuir para o desenvolvimento de métodos de inferência de redes genéticas e assim também acrescentar recursos ao trabalho de desenvolvimento métodos de diagnósticos e cura de doenças.

Capítulo 1

Introdução

Atualmente na área de Sistemas Biológicos trabalha-se para o desenvolvimento de ferramentas que possam auxiliar a obter um maior conhecimento sobre as interações moleculares em um organismo. Pesquisas tem sido desenvolvidas com o objetivo principal de obter um detalhamento da organização funcional de Sistemas Biológicos[Gibas and Jambeck, 2002, Lopes, 2011].

Estudos mostram que a alta complexidade dos organismos está intimamente relacionada com a organização de seus componentes em rede. Visando obter o detalhamento mencionado anteriormente, pesquisadores tem investido em estudos das redes biológicas. Desde que tenhamos à nossa disposição muitos tipos diferentes de medições a partir dos componentes de uma dessas redes, uma abordagem interessante seria tentar reconstruir essas redes. As redes regulatórias genéticas tem sido utilizadas como modelo para dispor mapas detalhados sobre as interações entre os genes e consequentemente sobre as interações moleculares e a organização funcional de organismos complexos. Porém, estas redes regulatórias são altamente complexas e o processo para inferi-las é computacionalmente custoso [Lopes, 2011]

As Redes Regulatórias Genéticas (GRN) são um conjunto de genes que interagem uns com os outros num organismo. Em uma GRN genes controlam a expressão de outros genes. Porém nem todos os genes da rede interagem entre si. Normalmente um gene controla apenas um subconjunto da rede e por sua vez é controlado por outro subconjunto de genes [Lopes, 2011].

Alguns dos métodos que tem sido utilizados para realizar a inferência de redes regu-

latórias genéticas são: Equações Diferenciais (ED) [Werhli, 2007a], Rede Bayesiana (BN) [Friedman et al., 2000], Redes Booleanas (BoN) [Shawn Martin et al., 2005], Redes Booleanas Probabilísticas (PBN) [Ilya Shmulevich and Zhang, 2002] entre outras.

Neste trabalho optou-se por usar redes Bayesianas devido à sua natureza probabilística e flexibilidade para incorporar intervenções e fontes adicionais de informações. Devido a, geralmente, os dados disponíveis serem esparsos e de ser impossível enumerar todas as redes existentes, as redes Bayesianas são amostradas através do BNMCMC que garante que haja uma convergência para distribuição posterior, desde que alguns requisitos sejam satisfeitos. Porém, na prática o BNMCMC é relativamente lento para alcançar a convergência e produzir os resultados esperados.

Ao mostrar-se lento em alguns casos, o BNMCMC exige muitas amostras para alcançar a distribuição posterior, de acordo com [Friedman and Koller, 2001]. Devido ao fato de em muitos dos casos os dados reais disponíveis não possuírem amostras abundantes, encontra-se um problema que pode dificultar a obtenção dos resultados. O conhecimento a priori utilizado na inferência de redes, geralmente, é proveniente de dados biológicos armazenados ou simulados. Neste trabalho ao invés de utilizar como conhecimento a priori dados biológicos ou simulados, utilizamos os resultados da inferência obtidos de outros métodos menos precisos, porém mais rápidos para direcionar o BNMCMC.

Visando não realizar um direcionamento estrito, o qual poderia violar as regras do BNMCMC, propõem-se então neste trabalho um direcionamento através de um modelo Hierárquico Bayesiano. Como método direcionador utilizamos o método GGM . Buscando avaliar os resultados obtidos através das simulações com o método BNGGM utilizamos MI também como conhecimento a priori. Assim, o objetivo principal deste trabalho ao propor o Modelo Hierárquico Bayesiano que utiliza o GGM ou MI como métodos direcionadores para o BNMCMC produzindo um melhoramento na convergência do método.

Nos capítulos a seguir apresentamos o desenvolvimento do trabalho, seus resultados e conclusões. No capítulo 2 é descrita a metodologia, no capítulo 3 são especificadas as configurações utilizadas na realização das simulações bem como os dados utilizados para realizar os experimentos, no capítulo 4 são apresentados os resultados das simulações realizadas com os métodos BNMCMC, BNGGM e BNMI. No capítulo 5 são apresentadas as conclusões e considerações finais com relação ao trabalho.

Capítulo 2

Metodologia

2.1 Redes Bayesianas (BN)

As BNs são uma combinação da teoria probabilística com a teoria dos grafos. Elas são muito úteis para representar relações probabilísticas entre várias entidades que interagem. Seus nós representam variáveis aleatórias e seus arcos representam dependências entre essas variáveis aleatórias. Formalmente uma BN é composta por uma estrutura gráfica $\mathcal{M}$, uma família de distribuições probabilísticas $\mathcal{F}$ e um conjunto de parâmetros q . Já $\mathcal{M}$ é composta por um conjunto de nós e arestas, onde os nós representam variáveis e as arestas representam as relações entre as variáveis. A estrutura $\mathcal{M}$ tem que ser um Grafo Acíclico Direcionado (DAG) [Adriano V. Werhli and Husmeier, 2006]. O conjunto de parâmetros q indica a natureza das interações entre os nós e a intensidade dessas interações [Werhli, 2007a].

A descrição da estrutura de uma rede gênica é feita por meio de BNs e é modelada pela DAG $\mathcal{M} = (V; E)$. Os vértices $v_i \in V$, $1 \leq i \leq n$ representam os n genes e correspondem a variáveis aleatórias X_i . Considerando que v_i é um gene, então X_i descreve o nível de expressão do gene v_i . Para cada X_i , uma distribuição de probabilidade condicional $\mathcal{P}(X_i | \text{preditores}(X_i))$ é definida, na qual $\text{preditores}(X_i)$ representam as variáveis que regulam diretamente o nodo v_i da rede $\mathcal{M}$ [Lopes, 2011].

No exemplo demonstrado na figura 2.1 tem-se a representação de uma rede Bayesiana composta por 10 nós e 11 arestas, a qual representa uma rede Bayesiana composta por :

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

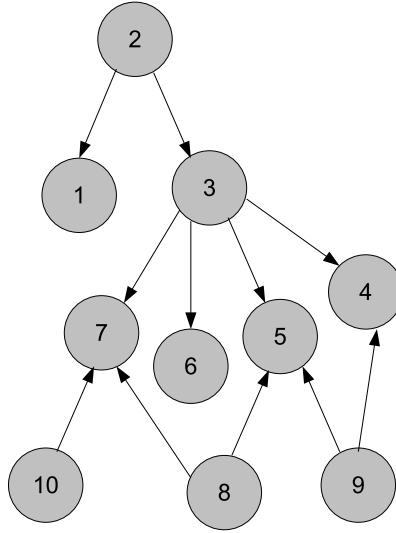


Figura 2.1: Exemplo de rede Bayesiana composta por 10 nós e 11 arestas

$E = \{(2, 1), (2, 3), (3, 7), (3, 6), (3, 5), (3, 4), (10, 7), (8, 7), (8, 5), (9, 5), (9, 4)\}$ onde V são os nós e E suas arestas.

Uma BN é caracterizada por uma regra simples para a expansão da probabilidade conjunta. Esta segue a propriedade de Markov: um nó é condicionalmente independente de seus não descendentes dado seus pais. Percebe-se então que são modelos estocásticos que incorporam aleatoriedade ou incerteza. Essa regra pode ser representada como:

$$\mathcal{P}(X_1, X_2, \dots, X_n) = \prod_{i=1}^n \mathcal{P}(X_i | \pi_{\mathcal{M}}(X_i)) \quad (2.1)$$

Ao aplicarmos a equação 2.1 podemos obter a probabilidade conjunta da rede representada na figura 2.1 que é definida por:

$$\begin{aligned} \mathcal{P}(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) &= \mathcal{P}(2) \mathcal{P}(1|2) \mathcal{P}(3|2) \mathcal{P}(7|3, 8, 10) \mathcal{P}(6|3) \\ &\quad \mathcal{P}(5|3, 8, 9) \mathcal{P}(4|3, 9) \mathcal{P}(8) \mathcal{P}(9) \mathcal{P}(10) \end{aligned} \quad (2.2)$$

Para especificar a distribuição conjunta é necessário determinar cada uma das probabilidades condicionais da equação 2.2 [Friedman et al., 2000]. E para representar as famílias $\mathcal{F}$ é possível escolher diferentes representações como por exemplo Gaussiana e Multinomial. Essa escolha dependerá do tipo variável que estamos lidando, podendo ser

variáveis discretas ou contínuas. Na representação linear Gaussiana é preciso discretizar os dados, mas só pode lidar com dependências que estão perto do linear. Já a representação multinomial pode capturar dependências que são não lineares.

Uma lacuna importante nas BNs é que não é possível modelar laços de realimentação que comumente estão presentes em redes biológicas reais. Devido a esse fato é necessário verificar a aciclidade das estruturas. Esta verificação é um dos gargalos das simulações MCMC. Um maneira de resolver esse problema é considerar Redes Bayesinas Dinâmicas (DBNs). Quando encontramos uma BN inválida então obtemos um grafo adequado, a chamada Rede Bayesina Dinâmica (DBN). Na prática precisamos apenas modificar a equação 2.1 a fim de incorporar o pressuposto de Markov em primeira ordem, o que implica que um $\mathcal{X}_i(t)$ em um tempo t tem pais $\pi_{\mathcal{M}}(\mathcal{X}_i)(t-1)$ em tempo $t-1$:

$$\mathcal{P}(X_1, X_2, \dots, X_n) = \prod_{i=1}^n \mathcal{P}(X_i(t) | X_{\pi_{\mathcal{M}}(\mathcal{X}_i)}(t-1)) \quad (2.3)$$

Para maiores detalhes sobre as DBNs consultar [Friedman et al., 1998].

2.1.1 Aprendizagem de BNs

Aprendizagem de uma BN significa que queremos ter uma DAG com um conjunto de probabilidades condicionais parametrizadas que melhor explanam os dados. Para aprender uma BN não é estritamente necessário usar Aprendizagem Bayesiana, mas neste trabalho utilizamos essa técnica. O processo de aprendizagem de BNs é realizado em duas etapas: primeiramente se aprende a estrutura, ou seja, as arestas que ligam as entidades e após aprende-se o conjunto de parâmetros associados as arestas e se o relacionamento entre essas arestas é de ativação ou inibição.

Se definirmos que $\mathbb{M}$ é o espaço de todos os modelos, o primeiro objetivo é encontrar o modelo $\mathcal{M}^* \in \mathbb{M}$ que é melhor suportado pelos dados $\mathcal{D}$, $\mathcal{M}^* = \operatorname{argmax}_{\mathcal{M}} \{P(\mathcal{M}|\mathcal{D})\}$. Tendo a melhor estrutura $\mathcal{M}^*$ e os dados $\mathcal{D}$, podemos encontrar os melhores parâmetros $\mathbf{q} = \operatorname{argmax}_{\mathbf{q}} \{P(\mathbf{q}|\mathcal{M}^*, \mathcal{D})\}$. Se aplicarmos a regra de Bayes temos:

$$P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{M})P(\mathcal{M}) \quad (2.4)$$

onde a probabilidade marginal implica uma integração ao longo de todo o espaço de

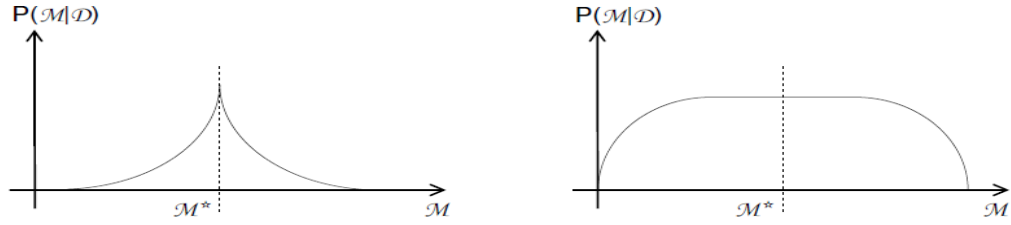


Figura 2.2: O eixo vertical mostra a probabilidade posterior $P(H|\mathcal{M})$ e o eixo horizontal representa a estrutura do modelo $\mathcal{M}$. O gráfico esquerdo mostra a probabilidade posterior de um conjunto de dados grande e informativo, onde a melhor estrutura $\mathcal{M}^*$ é muito bem definida. No gráfico da direita o conjunto de dados é pequeno e menos informativo e há muitas estruturas com alta probabilidade posterior de pontuação que levam a uma grande incerteza sobre a melhor estrutura.

parâmetros:

$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\mathbf{q}, \mathcal{M})P(\mathbf{q}|\mathcal{M})d\mathbf{q} \quad (2.5)$$

A integral na equação é analiticamente tratável quando os dados estão completos e o conhecimento a priori $P(\mathbf{q}|\mathcal{M})$ e a probabilidade $P(\mathcal{D}|\mathbf{q}, \mathcal{M})$ satisfazem certas condições [Heckerman, 1994, Heckerman, 1995].

No gráfico à direita da figura 2.2 podemos ver claramente a representação da grande quantidade de estruturas resultantes das inferências e assim percebemos a incerteza sobre a melhor estrutura para representar o modelo.

De acordo com a equação temos uma maneira de atribuir uma pontuação para uma estrutura gráfica dado um conjunto de dados. Porém encontrar a pontuação mais alta não é algo trivial. Um dos problemas associados a essa questão é a quantidade muito elevada do número de nós que compõem a estrutura. Além disso, quando temos um conjunto de dados esparsos, $P(\mathcal{M}|\mathcal{D})$ que é difuso, significando que $P(\mathcal{M}|\mathcal{D})$ não será representado adequadamente por uma estrutura $\mathcal{M}^*$. Por isso, um esquema MCMC é adotado [Madigan and York, 1995], que garante que sob algumas condições é teoricamente garantido que haja a convergência para distribuição posterior [Hastings, 1970].

A métrica de *score* utilizada para medição de dados discretizados é a métrica *Discrete*

Multinomial Bayesian scoring metric (BDe). Neste *score* assume-se que cada variável é associada com uma distribuição multinomial. Embora a discretização pode acarretar perda de informação, o *score* é bem flexível, pois pode modelar relações não-lineares.

Já para a medição de dados contínuos a métrica de *score* utilizada é *Continuous Bayesian Gaussian scoring metric* (BGe). Neste caso assume-se que cada variável está associada a uma distribuição Gaussiana.

2.2 Redes Bayesianas com MCMC (BNMCMC)

Constata-se que as BNs tem como principais vantagens o fato de permitirem trabalhar com incertezas e também por serem de fácil visualização e entendimento para não especialistas. E pelo fato de as BNs apresentarem um relacionamento probabilístico entre variáveis, são ferramentas muito utilizadas para o estudo de redes gênicas. Por esta razão, escolhemos esta ferramenta para o desenvolvimento deste trabalho.

Diante do fato mencionado é mais adequado amostrar as redes a partir da probabilidade posterior:

$$\mathcal{P}(\mathcal{M}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M})\mathcal{P}(\mathcal{M})}{\mathcal{P}(\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M})\mathcal{P}(\mathcal{M})}{\sum_{\mathcal{M}'} \mathcal{P}(\mathcal{D}|\mathcal{M}')\mathcal{P}(\mathcal{M}')} \quad (2.6)$$

Uma aproximação direta para encontrar $\mathcal{P}(\mathcal{M}|\mathcal{D})$ é impossível. A solução para esse problema é criar uma cadeia de Markov. E essa cadeia tem a seguinte forma:

$$\mathcal{P}_{n+1}(\mathcal{M}_{new}) = \sum_{old} \mathcal{T}(\mathcal{M}_{new}|\mathcal{M}_{old})\mathcal{P}_n(\mathcal{M}_{old}) \quad (2.7)$$

onde, $\mathcal{P}(\mathcal{M}_{new}) | (\mathcal{M}_{old})$ representa a matriz de transição de $(\mathcal{M}_{old})$ para $(\mathcal{M}_{new})$

$$\mathcal{P}_{\infty}(\mathcal{M}_{new}) = \sum_{old} \mathcal{T}(\mathcal{M}_{new}|\mathcal{M}_{old})\mathcal{P}_{\infty}(\mathcal{M}_{old}) \quad (2.8)$$

e o que precisa-se obter é :

$$\mathcal{P}(\mathcal{M}|\mathcal{D}) = \mathcal{P}_{\infty}(\mathcal{M}) \quad (2.9)$$

Ou seja, a cadeia de Markov convergindo para a probabilidade posterior. [Toni and Stumpf, 2010]

A equação de balanço para que essa condição seja suficiente é:

$$\frac{\mathcal{T}(\mathcal{M}_{new}|\mathcal{M}_{old})}{\mathcal{T}(\mathcal{M}_{old}|\mathcal{M}_{new})} = \frac{\mathcal{P}(\mathcal{M}_{new}|\mathcal{D})}{\mathcal{P}(\mathcal{M}_{old}|\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_{new})\mathcal{P}(\mathcal{M}_{new})}{\mathcal{P}(\mathcal{D}|\mathcal{M}_{old})\mathcal{P}(\mathcal{M}_{old})} \quad (2.10)$$

Sendo $Q(\mathcal{M}_{new}|\mathcal{M}_{old})$, a equação para as probabilidades de aceitação é a seguinte:

$$\frac{\mathcal{M}(\mathcal{M}_{new}|\mathcal{M}_{old})}{\mathcal{A}(\mathcal{M}_{old}|\mathcal{M}_{new})} = \frac{\mathcal{P}(\mathcal{M}_{new}|\mathcal{D})}{\mathcal{P}(\mathcal{M}_{old}|\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_{new})\mathcal{P}(\mathcal{M}_{new})Q(\mathcal{M}_{old}|\mathcal{M}_{new})}{\mathcal{P}(\mathcal{D}|\mathcal{M}_{old})\mathcal{P}(\mathcal{M}_{old})Q(\mathcal{M}_{new}|\mathcal{M}_{old})} \quad (2.11)$$

A condição suficiente é:

$$\mathcal{A}(\mathcal{M}_{new}|\mathcal{M}_{old}) = \min \left\{ \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_{new})\mathcal{P}(\mathcal{M}_{new})Q(\mathcal{M}_{old}|\mathcal{M}_{new})}{\mathcal{P}(\mathcal{D}|\mathcal{M}_{old})\mathcal{P}(\mathcal{M}_{old})Q(\mathcal{M}_{new}|\mathcal{M}_{old})}, 1 \right\} \quad (2.12)$$

A partir dessas derivações anteriores, percebemos que aceitar uma nova configuração $\mathcal{M}_{new}$ com probabilidade dada pela equação 2.12 é a condição para satisfazer a equação de equilíbrio detalhado em 2.10, que garante que a cadeia de Markov irá convergir para a distribuição desejada da equação 2.6.

Esta equação pode ser reescrita da seguinte forma:

$$\mathcal{A}(\mathcal{M}_{new}|\mathcal{M}_{old}) = \min \left\{ \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_{new})\mathcal{P}(\mathcal{M}_{new})}{\mathcal{P}(\mathcal{D}|\mathcal{M}_{old})\mathcal{P}(\mathcal{M}_{old})}, 1 \right\} \quad (2.13)$$

pois assim obtemos a distribuição proposta simétrica.

A equação (2.12) é conhecida como critério de aceitação Metropolis-Hastings.

Neste trabalho utilizamos as BNs com MCMC afim de realizar a inferência das redes genéticas em estudo. A figura 2.3 mostra o processo do MCMC já descrito e na figura 2.4 temos a representação do processo das simulações BNMCMC.

2.3 Combinando Conhecimento a Priori com a Expressão Gênica

O objetivo deste trabalho é estudar a integração do conhecimento biológico prévio para a inferência de redes regulatórias genéticas. Para realizar as inferências utilizando conhecimento a priori propomos o modelo hierárquico Bayesiano apresentado na figura 2.5. Neste

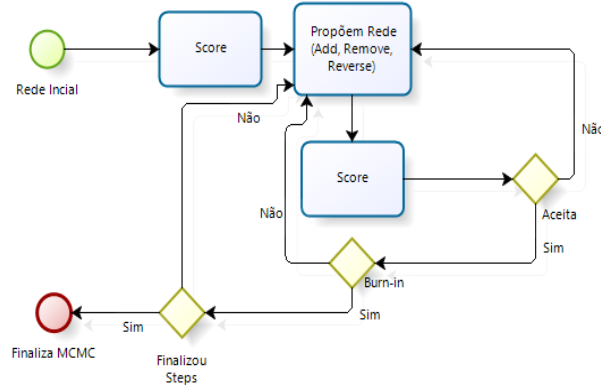


Figura 2.3: Representação das etapas do processo do MCMC

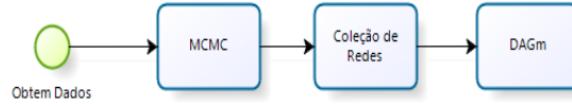


Figura 2.4: Representação do processo das simulações do modelo BNMCMC

modelo é necessário medir a diferença entre o modelo $\mathcal{M}$ e o conhecimento a priori que dispomos, por exemplo, considerando a matriz M representada na equação 2.14 como sendo o modelo $\mathcal{M}$ e a matriz P representada na equação 2.15 como a matriz de conhecimento a priori, para ter o valor dessa energia mede-se o quanto a matriz M difere da matriz P . Para efetuar essa medida seguimos a aproximação proposta por [Imoto et al., 2003] que é chamada de medida de energia $\mathcal{E}$. Onde $\mathcal{E}$ é definida pela equação 2.17.

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (2.14)$$

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (2.15)$$

A rede $\mathcal{M}$ é a matriz de adjacência. Nessa matriz a entrada $\mathcal{M}_{ij}$ pode ser 0 ou 1, representando a presença ou ausência de uma aresta entre dois nós i e j . É definida como

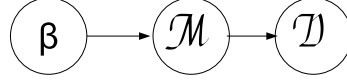


Figura 2.5: Este esquema apresenta o Modelo Gráfico Probabilístico, onde temos o conhecimento priori que representa o conhecimento das interações entre os nós e então é realizada a simulação da Rede Bayesiana, assim é gerado o modelo $\mathcal{M}$ resultante da simulação que é então compara com a rede verdadeira dos dados $\mathcal{D}$

matriz de conhecimento à priori a matriz que contém como entrada i e j , que representa o conhecimento das interações entre os nós. Se $\mathcal{B}_{ij} = 0.5$ não há conhecimento sobre as relações entre os nós, se $0 < \mathcal{B}_{ij} < 0.5$ há evidência prévia de que não há nenhuma aresta direcionada entre os nós i e j , se $0.5 < \mathcal{B}_{ij} \leq 1$ há evidência prévia de que há uma aresta direcionada entre os nós i e j .

A equação 2.16 representa o escalonamento da matriz de correlação parcial.

$$\rho_{sik} = \frac{|\rho_{ik}| - \min(\boldsymbol{\rho})}{\max(\boldsymbol{\rho}) - \min(\boldsymbol{\rho})}, \quad (2.16)$$

onde $\{\rho_{ik} \in \mathbb{R} | -1 \leq \rho_{ik} \leq 1\}$.

A Energia mencionada anteriormente é a medida de concordância entre a rede $\mathcal{M}$ e o conhecimento prévio disponibilizado através de dados biológicos ou através do resultados de simulações. A energia pode ser obtida da seguinte maneira:

$$\mathcal{E}(\mathcal{M}) = \sum_{i,j=1}^N |\rho_{s_{i,j}} - \mathcal{M}_{i,j}| \quad (2.17)$$

Onde N é o número total de nós. A energia $\mathcal{E}(\mathcal{M}) = 0$ quando ocorre uma combinação perfeita entre a rede $\mathcal{M}$ e o conhecimento prévio $\mathcal{B}$ e os valores crescentes de $\mathcal{E}$ indicam um descompasso crescente entre $\mathcal{M}$ e $\mathcal{B}$.

Conhecendo a distribuição a priori no hiperparametro β , $\mathcal{P}(\beta)$, o objetivo agora é amostrar a estrutura da rede $\mathcal{M}$ e o hiperparametro β da distribuição posteriori $\mathcal{P}(\mathcal{M}|\beta, \mathcal{D})$. Para este fim uma nova estrutura de rede $\mathcal{M}_{new}$ a partir da distribuição proposta $\mathcal{Q}(\mathcal{M}_{new}|\mathcal{M}_{old})$ e um novo hiperparametro da distribuição proposta $\mathcal{R}(\beta_{new}|\beta_{old})$. Aceita-se então os movimentos na estrutura da rede de acordo com padrão de regras do

Metropolis-Hastings [Hastings, 1970]. Com a seguinte probabilidade de aceitação:

$$A = \min \left\{ \frac{P(\mathcal{D}, \mathcal{M}_{new}, \beta_{new})Q(\mathcal{M}_{old}|\mathcal{M}_{new})R(\beta_{old}|\beta_{new})}{P(\mathcal{D}, \mathcal{M}_{old}, \beta_{old})Q(\mathcal{M}_{new}|\mathcal{M}_{old})R(\beta_{new}|\beta_{old})}, 1 \right\} \quad (2.18)$$

que, devido às relações de independência condicional representada pela figura 2.18 pode ser expandido da forma abaixo:

$$A = \min \left\{ \frac{P(\mathcal{D}|\mathcal{M}_{new})P(\mathcal{M}_{new}|\beta_{new})P(\beta_{new})Q(\mathcal{M}_{old}|\mathcal{M}_{new})R(\beta_{old}|\beta_{new})}{P(\mathcal{D}|\mathcal{M}_{old})P(\mathcal{M}_{old}|\beta_{old})P(\beta_{old})Q(\mathcal{M}_{new}|\mathcal{M}_{old})R(\beta_{new}|\beta_{old})}, 1 \right\} \quad (2.19)$$

Para aumentar a probabilidade de aceitação e maximizar a convergência da cadeia de Markov é admitido quebrar o movimento em dois submovimentos. Primeiramente amostramos uma nova estrutura de rede $\mathcal{M}_{new}$ para a distribuição proposta $\mathcal{Q}(\mathcal{M}_{new}|\mathcal{M}_{old})$, mantendo o hiperparâmetro β fixo e aceitando esse movimento através da seguinte probabilidade de aceitação:

$$A(\mathcal{M}_{new}|\mathcal{M}_{old}) = \min \left\{ \frac{P(\mathcal{D}|\mathcal{M}_{new})P(\mathcal{M}_{new}|\beta)Q(\mathcal{M}_{old}|\mathcal{M}_{new})}{P(\mathcal{D}|\mathcal{M}_{old})P(\mathcal{M}_{old}|\beta)Q(\mathcal{M}_{new}|\mathcal{M}_{old})}, 1 \right\} \quad (2.20)$$

Após amostramos um novo hiperparâmetro β da distribuição proposta $\mathcal{R}(\beta_{new}|\beta_{old})$ para um estrutura de rede fixa $\mathcal{M}$ e aceitamos o movimento através da probabilidade de aceitação abaixo:

$$A(\beta_{new}|\beta_{old}) = \min \left\{ \frac{P(\mathcal{M}_{old}|\beta_{new})P(\beta_{new})R(\beta_{old}|\beta_{new})}{P(\mathcal{M}_{old}|\beta_{old})P(\beta_{old})R(\beta_{new}|\beta_{old})}, 1 \right\} \quad (2.21)$$

Para uma distribuição a priori uniforme $\mathcal{P}(\beta)$ e uma distribuição proposta simétrica $\mathcal{R}(\beta_{new}|\beta_{old})$ temos a expressão simplificada a seguir:

$$A(\beta_{new}|\beta_{old}) = \min \left\{ \frac{P(\mathcal{M}_{old}|\beta_{new})}{P(\mathcal{M}_{old}|\beta_{old})}, 1 \right\} \quad (2.22)$$

Então esses procedimentos são iterados até que algum critério de convergência seja satisfeito. Neste trabalho estamos usando um critério de convergência baseado no método heurístico visual, este método de convergência que propomos está descrito na subseção 3.3.2. A partir do momento que este critério de convergência é aceito para-se as iterações. A equação da probabilidade de aceitação pode ser reescrita como:

$$A(\beta_{new}|\beta_{old}) = \min \left\{ \frac{e^{-\mathcal{E}(\mathcal{M})(\beta_{new}-\beta_{old})} Z(\beta_{old})}{Z(\beta_{new})}, 1 \right\} \quad (2.23)$$

Essa equação mostra a dependência da probabilidade de aceitação nas funções de distribuição $Z(\beta_{old})$ e $Z(\beta_{new})$. O cálculo das funções de distribuição implica um somatório ao longo de todo o espaço das estruturas de rede que, devido à sua complexidade super-exponencial é impraticável obter. No entanto, considerando que todas as redes possíveis são válidas, podemos reduzir esta complexidade polinomial, tornando-se assim possível obter um limite superior sobre a função de distribuição verdadeira. Para uma discussão detalhada sobre este assunto, ver [Werhli, 2007b, Werhli and Husmeier, 2007]. Escolhemos a distribuição a priori do hiperparâmetro $\mathcal{P}(\beta)$ para ser a distribuição sobre o intervalo de $[0, \text{MAX}]$

A figura 2.5 representa o funcionamento do Modelo Probabilístico, onde são utilizados conhecimentos a priori para direcionar o MCMC.

2.4 Graphical Gaussian Models (GGM)

Graphical Gaussian Models (GGMs) são modelos gráficos probabilísticos não direcionados que identificam as relações de independência condicional entre os nós sob a suposição de uma distribuição Gaussiana multivariada dos dados. A inferência de GGMs é baseada na estimação da matriz de covariância dessa distribuição [Werhli, 2007a]. O elemento C_{ik} da matriz de covariância $\mathbf{C}$ está relacionado com o coeficiente de correlação entre os nós X_i and X_k . Um coeficiente de correlação alto entre dois nós pode indicar uma interação direta, uma interação indireta ou uma regulação conjunta por um fator em comum. Porém, somente interações diretas interessam para a construção de uma rede regulatória. A força dessas interações é medida pelo coeficiente de correlação parcial ρ_{ik} que descreve a correlação entre os nós X_i and X_k . Da teoria de distribuição normal, veja [Edwards, 2000], sabe-se que a matriz $\boldsymbol{\rho}$ de coeficientes de correlação parcial ρ_{ik} está relacionada com a inversa da matriz de covariância $\mathbf{C}$, $\mathbf{C}^{-1}$:

$$\rho_{ik} = - \left(\frac{C_{ik}^{-1}}{\sqrt{C_{ii}^{-1} C_{kk}^{-1}}} \right) \quad (2.24)$$

Comumente, no momento de inferir um GGM, a matriz de covariância é computada empiricamente, invertida e as correlações parciais ρ_{ik} são computadas através da equação (2.24). Na verificação da distribuição de $|\rho_{ik}|$ as arestas (i, k) que correspondem a valores significativamente pequenos são removidos do grafo. O passo crítico nesse procedimento é a estimativa estável da matriz de covariância e sua inversa.

Em [Schaefer and Strimmer, 2005] os autores propõem um estimador de covariância regularizado singular que está baseado no conceito de *shrinkage* e explora o lema de Ledoit Wolf [Ledoit and Wolf, 2004] para realizar o cálculo analítico.

Esse estimador para a matriz de covariância é que garante que ela seja não singular, de modo que ela possa ser invertida para obter-se um novo estimador $\hat{\boldsymbol{\rho}} = (\hat{\mathbf{C}})^{-1}$ para a matriz $\boldsymbol{\rho}$. O novo estimador é baseado na seguinte idéia: sabe-se que o método da máxima probabilidade irrestrita $\hat{\mathbf{C}}_{\text{ML}}$ para a matriz de covariância $\mathbf{C}$ que tem uma elevada variação, se o número de variáveis excede o número de observações. Por outro lado, existem muitos outros possíveis estimadores que possuem uma certa predisposição, mas uma variância baixa. A abordagem do *shrinkage* combina um estimador de verossimilhança máxima com um desses outros estimadores $\hat{\mathbf{C}}_T$ em uma média ponderada:

$$\hat{\mathbf{C}} = (1 - \lambda) \cdot \hat{\mathbf{C}}_{\text{ML}} + \lambda \cdot \hat{\mathbf{C}}_T, \quad (2.25)$$

onde $\lambda \in [0, 1]$ denota a intensidade da redução.

Os autores apresentam uma variedade de possíveis alvos da matriz de covariância, $\hat{\mathbf{C}}_T$. No entanto, eles recomendam e utilizam em seus experimentos a "diagonal, variância desigual" como estimador. Esta meta é definida na equação 2.26.

Uma propriedade muito interessante e útil sobre o alvo escolhido é que, quando combinado com o estimador de verossimilhança máxima irrestrita, de acordo com a equação (2.25) a matriz é definida e positiva. Os elementos desta meta específica são dadas por :

$$\hat{C}_{T_{ij}} = \begin{cases} \hat{C}_{\text{ML}_{ii}} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (2.26)$$

e para a matriz de covariância alvo é mostrado que o estimador de *shrinkage* ótimo, $\hat{\lambda}^*$, é dado por

$$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(\hat{C}_{\text{ML}_{ij}})}{\sum_{i \neq j} \hat{C}_{\text{ML}_{ij}}^2} \quad (2.27)$$

onde $\sum_{i \neq j} \widehat{\text{Var}}(\hat{C}_{\text{ML}_{ij}})$ e $\hat{C}_{\text{ML}_{ii}}$ são estimativas imparciais obtidas a partir dos dados.

2.5 Informação Mútua

A informação mútua $I(X; Y)$ é a entropia relativa entre a distribuição conjunta e o produto das probabilidades marginais, como representado na equação (2.28). De maneira informal, é uma medida da quantidade de informação que uma variável aleatória contém acerca da outra. Sendo duas variáveis aleatórias X e Y com distribuição conjunta $p(x, y)$ e distribuições marginais $p(x)$ e $p(y)$.

$$I(X, Y) = H(X, Y) - H(X) - H(Y) \quad (2.28)$$

Onde a entropia conjunta $H(X, Y)$ de X e Y é representado pela equação (2.29)

$$H(X, Y) = - \sum_X \sum_Y p(x, y) \log(p(x, y)) \quad (2.29)$$

A entropia da variável aleatória X é representada pela equação (2.30)

$$H(X) = - \sum_X p(x) \log(p(x)) \quad (2.30)$$

A entropia da variável aleatória Y é representada pela equação (2.31)

$$H(Y) = - \sum_Y p(y) \log(p(y)) \quad (2.31)$$

A medida de informação mútua expressa a quantidade de informação que X compartilha com Y . Sendo assim, quando X e Y são independentes temos que $H(X|Y) = H(X)$ e $I(X; Y) = 0$. Se Y for função de X então $I(X; Y) = H(X) = H(Y)$ [Artuso, 2011, Paninski, 2003].

Neste trabalho utilizamos MI, em um segundo método, como geradora de conhecimento à priori em substituição do método GGM afim de realizar comparações com os resultados

de ambos os experimentos. Para cada nodo da rede em análise foi calculado o valor da informação mútua entre o nodo e seus possíveis pais.

2.6 BNs Guiadas por GGMs

Um dos maiores problemas na inferência das BNs ao utilizar MCMC é a demora para obter a convergência do MCMC. Assim, neste trabalho, propomos a utilização dos resultados da inferência GGM como auxílio para o algoritmo BN-MCMC que infere a BN. É proposto então um Modelo Bayesiano Hierárquico (BNGGM). Isto permite que a amostragem MCMC seja guiada por um método mais rápido. A figura 2.6 ilustra o funcionamento do Modelo Bayesiano Hierárquico. O Modelo Gráfico Probabilístico representa as relações de independência condicional entre os dados $\mathcal{D}$, o modelo da rede $\mathcal{M}$ e o hiperparâmetro β_{GGM} associado ao GGM. O resultado da inferência GGM é a matriz de adjacência, ρ_s , onde ρ_{ij} é o coeficiente de correlação parcial entre as variáveis i e j [Werhli and Husmeier, 2007].

Segundo [Werhli and Husmeier, 2007] a distribuição prévia toma a forma da distribuição de Gibbs representada pela equação 2.32

$$\mathcal{P}(\mathcal{M}|\beta_{GGM}) = \frac{e^{-\beta_{GGM}\mathcal{E}(\mathcal{M})}}{Z(\beta_{GGM})} \quad (2.32)$$

Onde β_{GGM} é o hiperparâmetro associado à GGM, $Z(\beta_{GGM})$ é a constante de normalização e $\mathcal{E}(\mathcal{M})$ é a energia da rede $\mathcal{M}$. A energia de uma rede é definida como a medida de quanto as redes são similares. O que é proposto neste trabalho é que sejam amostradas as redes que são mais similares às redes geradas como resultados das simulações GGMs. O hiperparametro β pode ser interpretado como um fator que indica a força da influência do conhecimento prévio biológico com relação aos dados. Para $\beta \rightarrow 0$ a distribuição a priori torna-se pouco informativa sobre a estrutura de rede. Porém, para $\beta \rightarrow \infty$ a distribuição prévia torna-se informativa referente a estrutura que atingiu a energia mais baixa.

O objetivo da inferência de BNs guiadas por GGMs é amostrar a estrutura de rede de $\mathcal{M}$ e o hiperparâmetro β_{GGM} da distribuição posterior $P(\mathcal{M}, \beta_{GGM}|\mathcal{D})$. Para isso uma nova estrutura de rede de $(\mathcal{M}')$ e um novo hiperparâmetro (β'_{GGM}) são propostos, respec-

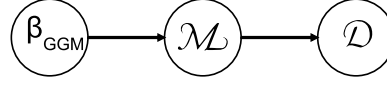


Figura 2.6: Este esquema apresenta o Modelo Bayesiano Hierárquico (BNGGM), onde temos o conhecimento a priori GGM que representa o conhecimento das interações entre os nós e então é realizada a simulação da BN utilizando o GGM como conhecimento a priori do BNMCMC, assim é gerado o modelo $\mathcal{M}$ resultante da simulação que é então compara com a rede verdadeira dos dados $\mathcal{D}$

tivamente a partir das propostas distribuições $Q(\mathcal{M}'|\mathcal{M})$ e $R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})$.

$$A = \min \left\{ \frac{P(\mathcal{D}, \mathcal{M}', \beta'_{\text{GGM}})Q(\mathcal{M}|\mathcal{M}')R(\beta_{\text{GGM}}|\beta'_{\text{GGM}})}{P(\mathcal{D}, \mathcal{M}, \beta_{\text{GGM}})Q(\mathcal{M}'|\mathcal{M})R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})}, 1 \right\} \quad (2.33)$$

devido a relação de independência condicional descrita na figura 2.6 pode ser expandida como segue:

$$A = \min \left\{ \frac{P(\mathcal{D}|\mathcal{M}')P(\mathcal{M}'|\beta'_{\text{GGM}})P(\beta'_{\text{GGM}})Q(\mathcal{M}|\mathcal{M}')R(\beta_{\text{GGM}}|\beta'_{\text{GGM}})}{P(\mathcal{D}|\mathcal{M})P(\mathcal{M}|\beta_{\text{GGM}})P(\beta_{\text{GGM}})Q(\mathcal{M}'|\mathcal{M})R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})}, 1 \right\} \quad (2.34)$$

Para obter a amostragem da estrutura $\mathcal{M}$, primeiramente amostramos a novo modelo $\mathcal{M}'$ a partir da distribuição $Q(\mathcal{M}'|\mathcal{M})$, mantendo o hiperparâmetro β_{GGM} fixo e aceita-se o movimento com seguinte probabilidade de aceitação.

$$A(\mathcal{M}'|\mathcal{M}) = \min \left\{ \frac{P(\mathcal{D}|\mathcal{M}')P(\mathcal{M}'|\beta_{\text{GGM}})Q(\mathcal{M}|\mathcal{M}')}{P(\mathcal{D}|\mathcal{M})P(\mathcal{M}|\beta_{\text{GGM}})Q(\mathcal{M}'|\mathcal{M})}, 1 \right\} \quad (2.35)$$

Após, amostramos o novo hiperparâmetro β_{GGM} da distribuição proposta $R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})$ para uma estrutura de rede fixa $\mathcal{M}$ e aceita-se através probabilidade a seguir:

$$A(\beta'_{\text{GGM}}|\beta_{\text{GGM}}) = \min \left\{ \frac{P(\mathcal{M}|\beta'_{\text{GGM}})P(\beta'_{\text{GGM}})R(\beta_{\text{GGM}}|\beta'_{\text{GGM}})}{P(\mathcal{M}|\beta_{\text{GGM}})P(\beta_{\text{GGM}})R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})}, 1 \right\} \quad (2.36)$$

Para obter uma distribuição a priori uniforme $P(\beta_{\text{GGM}})$ e uma distribuição proposta simétrica $R(\beta'_{\text{GGM}}|\beta_{\text{GGM}})$, simplifica-se a expressão e obtem-se:

$$A(\beta'_{\text{GGM}}|\beta_{\text{GGM}}) = \min \left\{ \frac{P(\mathcal{M}|\beta'_{\text{GGM}})}{P(\mathcal{M}|\beta_{\text{GGM}})}, 1 \right\} \quad (2.37)$$

Esses dois passos são iterados até que algum critério de convergência seja satisfeito. A equação da probabilidade de aceitação pode ser reescrita como:

$$A(\beta'_{\text{GGM}}|\beta_{\text{GGM}}) = \min \left\{ \frac{e^{-E(\mathcal{M})(\beta'_{\text{GGM}} - \beta_{\text{GGM}})} Z(\beta_{\text{GGM}})}{Z(\beta'_{\text{GGM}})}, 1 \right\} \quad (2.38)$$

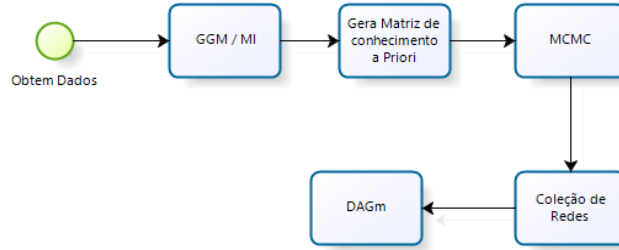


Figura 2.7: Representação do processo das simulações do modelo BNGGM / BNMI

Mediante o desenvolvimento deste processo é possível realizar a inferência de BNs através do BNMCML guiado por GGM. Este modelo é representado na figura 2.7.

2.7 BNs Guiadas por MI

Propomos também neste trabalho a utilização dos resultados da Informação Mútua como auxílio para o algoritmo BNMCML que infere a BN. É proposto então um BNMI. Isto também permite que a amostragem MCMC seja guiada por um método mais rápido. A figura 2.6 que ilustra o funcionamento do BNGGM pode representar o funcionamento do BNMI.

O procedimento realizado para a inferência de BNs guiadas por MI é basicamente o mesmo da inferência de BNs guiadas por GGM, a principal diferença entre eles é que no primeiro utilizamos o GGM como conhecimento a priori e neste último usamos Informação Mútua como método direcionador do BNMCML. Este modelo é representado na figura 2.7.

Capítulo 3

Simulações

Neste trabalho utilizamos três tipos de dados: dados gerados pela ferramenta GNW, dados Gaussianos e dados Reais. Os dados que utilizamos são descritos a seguir.

3.1 Dados

3.1.1 Dados GeneNetWeaver - Ecoli

Utilizou-se a ferramenta GNW [Thomas Schaffter and Roulet, 2011] desenvolvida na linguagem Java pelo grupo dos criadores do DREAM challenge [Thomas Schaffter and Roulet, 2006]. Para gerar os conjuntos de dados da rede representada pela figura 3.1 inicialmente é necessário gerar o modelo cinético através da opção *Generate Kinetic Model*. Esta opção cria um modelo dinâmico da rede Genética que será necessário para gerar os conjuntos de dados. Após ter gerado o modelo dinâmico pode-se efetuar a configuração de parâmetros. Foi utilizado para a geração dos dados o modelo *Deterministic (ODEs)*, ou seja, os dados são gerados a partir de um sistema estocástico de Equações Diferenciais Acopladas (EDA) com ruído adicionado. Este tipo de dados é mais parecido com dados reais, uma vez que apresenta não-linearidades que são típicas de sistemas biológicos reais. No processo de geração selecionamos experimentos para ser “*multifactorial*” com “*add Gaussian noise*” and “*std dev = 0.005*”. Os dados gerados por esse método a partir daqui em diante serão chamados de Ecoli.

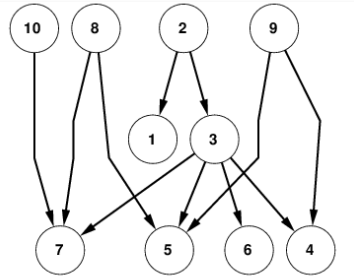


Figura 3.1: Rede - GNW: O grafo mostra a rede gerada pelo GNW, nós representam os genes e arestas representam as interações.

3.1.2 Dados Gaussianos de distribuição Gaussiana Multivariável

Uma maneira simples e clara de gerar dados sintéticos de uma estrutura é a amostragem de uma distribuição linear Gaussiana. A variável randômica X_i denota a expressão de um nó i que está distribuído de acordo com $X_i \sim N(\sum_k w_{ik}x_k, \sigma^2)$ onde $N(\cdot)$ denota a distribuição normal, o somatório se estende sobre todos os pais do nó i , e x_k representa o valor do nó k . A interação entre os nós X_i e X_k é $w_{ik} \neq 0$. Se $w_{ik} = 0$ então o nó X_k não é um pai do nó X_i . O valor de σ^2 pode ser interpretado como sendo um ruído da dinâmica. Valores baixos do σ^2 indicam um conjunto de dados muito determinístico, ou seja, o valor do nó filho é completamente determinado pelos valores dos nós pais. Contrariamente, valores altos de σ^2 indicam valores altos de ruídos nos dados. Esse processo é equivalente a amostragem de uma distribuição Gaussiana Multivariada. Para a geração de dados de uma estrutura de rede é necessário a ordenação topológica dos nós. Para geração dos dados Gaussianos foi utilizada como rede base a mesma rede gerada pelos conjuntos de dados obtidos da ferramenta GNW. Também foram gerados 5 conjuntos de dados com 100 amostras cada um.

3.1.3 Dados Reais

Experimentos de citometria de fluxo são utilizados para medição dos níveis de concentração de proteínas que compõem uma rede. Em [Sachs et al., 2005] os autores usaram experimentos de citometria de fluxo para medir os níveis de concentração de 11 proteínas que compõem a rede representada na figura 3.2. Essa rede está envolvida na proliferação

celular em células do sistema imunológico humano. A desregulação dessa rede pode causar carcinogênese (formação do câncer). Assim, esta rede é amplamente estudada como uma rede padrão obtida a partir de vários estudos distintos. Os dados produzidos através desse método são adotados como dados reais utilizados neste trabalho.

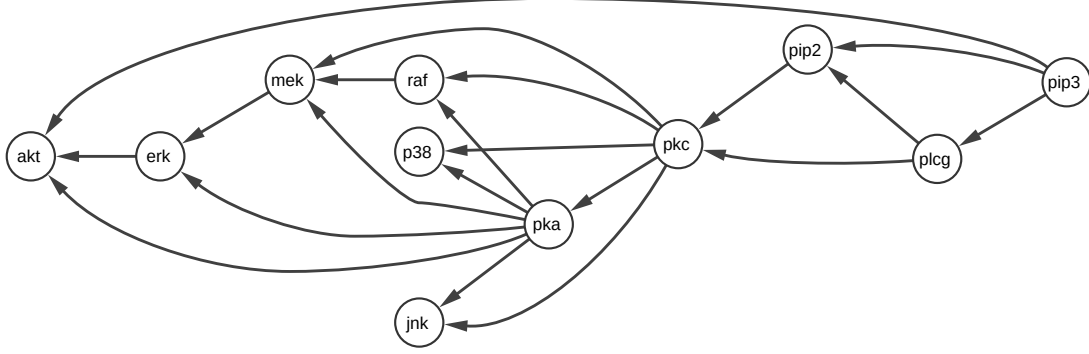


Figura 3.2: Rede - experimento citométrico: O grafo mostra a rede, nós representam as proteínas e arestas representam as interações.

3.2 Critérios de avaliação

O resultado das simulações é uma coleção de redes representadas pela matriz de adjacência. Dessa coleção de matrizes podemos obter uma matriz média, $\mathcal{R}$, onde cada entrada r_{ij} indica a ocorrência média de cada aresta na coleção de redes inseridas. Como forma de avaliar o desempenho das simulações é necessário comparar seu resultado com alguma rede conhecida. Chamamos essa rede conhecida de rede verdadeira, $\mathcal{T}$, onde as entradas $t_{ij} \in \{0, 1\}$, indicam a presença e a ausência de conexões entre os nodos X_i and X_j .

A fim de comparar a rede $\mathcal{R}$ com a rede verdadeira $\mathcal{T}$, podemos transformá-la numa matriz de adjacência, $\mathcal{A}_{\mathcal{R}}(\epsilon)$, através da imposição de um limite ϵ . Cada entrada da matriz de adjacência, a_{ij} , é 1 se $r_{ij} \geq \epsilon$ e 0 caso contrário.

Tendo essas duas matrizes, $\mathcal{T}$ e $\mathcal{A}_{\mathcal{R}}(\epsilon)$, podemos classificar cada uma das arestas em categorias. Uma aresta pode ser classificada como: Verdadeira Positiva (TP), Falsa Positiva (FP), Verdadeira Negativa (TN) ou Falsa Negativa (FN). A tabela 3.1 mostra um resumo de como as arestas são classificadas nessas categorias.

A ROC é obtida através da variação do limiar ϵ e plotando o número relativo de arestas

t_{ij}	a_{ij}	Categoria
0	0	TN
0	1	FP
1	0	FN
1	1	TP

Tabela 3.1: Classificação de arestas: Essa tabela mostra como uma aresta é classificada de acordo com os valores na matriz verdadeira (t_{ij}) e na matriz de adjacência (a_{ij}). Em uma entrada que é igual a zero significa que a aresta do nodo X_i para o nodo X_j está ausente, inversamente, uma entrada que é igual a um significa que a aresta está presente.

TP contra o número de arestas FP para cada um dos limiares. O ideal seria comparar toda a curva ROC, mas isso é impraticável. Então, usamos a Área abaixo da ROC (AUC) como resumo dos resultados para todos os limiares. Um preditor perfeito produziria um valor de AUC de 1.00. Por outro lado, um preditor aleatório produziria um valor de AUC em torno de 0.50. Geralmente, valores de área maiores representam melhores preditores.

3.3 Critérios de convergência

Usualmente, a fim de realizar diagnósticos de convergência, são analisados gráficos ou medidas descritivas. Métodos descritos em [Cowles and Carlin, 2006] ; Garren and Smith (2003); Gewke (1992); Johnson (1994); Liu, Liu and Rubin (1992); Mykland, Tierney and Yu (1995); Ritter and Tanner (1992); Roberts (1992, 1994); Yu (1994); Yu and Mykland (1994); and Zellner and Min; and also Heidelberge and Welch (1983) são utilizados para diagnosticar a convergência. A simulação de Cadeia de Markov é uma ferramenta poderosa e fácil de aplicar, porém contém sérios riscos de erro, incluindo modelagem inapropriada, erros em cálculos ou programação e convergência lenta [Gelman, 1996]. Para utilizar o BNMC MC é necessário determinar quanto tempo a simulação precisa ser executada, [Raftery and Lewis, 1995]. Então para obter o nível de precisão dos algoritmos BNMC MC é necessário fazer alguns testes de convergência. Esses testes podem ser efetuados de forma gráfica e/ou através dos métodos descritos anteriormente.

A maioria dos métodos é desenvolvida para problemas em que são amostrados na

cadeia de Markov, normalmente, no máximo três parâmetros. Problemas onde os diagnósticos de convergência amostram vários parâmetros simultaneamente são raros. A seguir é descrito e explanado um novo método para verificar a indicação de convergência no caso de amostradores BNMC MC de alta dimensionalidade.

3.3.1 Avaliação visual Heurística

Na amostragem de uma estrutura de rede cada amostra é uma matriz de adjacência. Para uma rede com n nodos a matriz de adjacência tem dimensão $n \times n$. Então, neste caso, temos que avaliar a convergência de n^2 parâmetros que são probabilidades posteriores sobre as arestas da rede. Nesse cenário, o tempo de convergência das simulações BNMC MC é alto. Assumimos que a matriz de probabilidades posterior é $\mathcal{P}$ e cada entrada p_{ij} da matriz indica a probabilidade posterior marginal da existência de uma aresta entre os nodos i e j . Para acessar a convergência do BNMC MC duas simulações são executadas com inicializações diferentes, obtendo como resultado duas matrizes de probabilidade posterior, $\mathcal{P}^1$ e $\mathcal{P}^2$. Após um gráfico de dispersão é produzido através da representação gráfica p_{ij}^1 contra p_{ij}^2 , para $1 < i, j < n$ [?]. A figura 3.3 exemplifica como são mostrados os resultados das simulações através dos gráficos de dispersão. Este método exige então uma análise visual dos resultados mostrados para avaliar a convergência que pode não ser alcançada algumas vezes.

3.3.2 O método proposto de avaliação de convergência

O método proposto está baseado no método heurístico visual apresentado na subseção 3.3.1. Ao invés de realizar a análise visual, propomos calcular o valor numérico que resume o que é visualizado. Considerando que cada passo do BNMC MC do parâmetro amostrado é uma matriz de parâmetros $\mathcal{P}$ com dimensão $n \times n$. As entradas da matriz são p_{ij} , $1 < i, j < n$. Primeiramente calculamos a distância de todos os pontos até a linha $y = x$. A distância de um ponto $P(x_i, y_i)$ para a linha representada pela equação $ax + by + c = 0$ é definida por:

$$d_i = \frac{|ax_i + by_i + c|}{\sqrt{(a^2 + b^2)}} \quad (3.1)$$

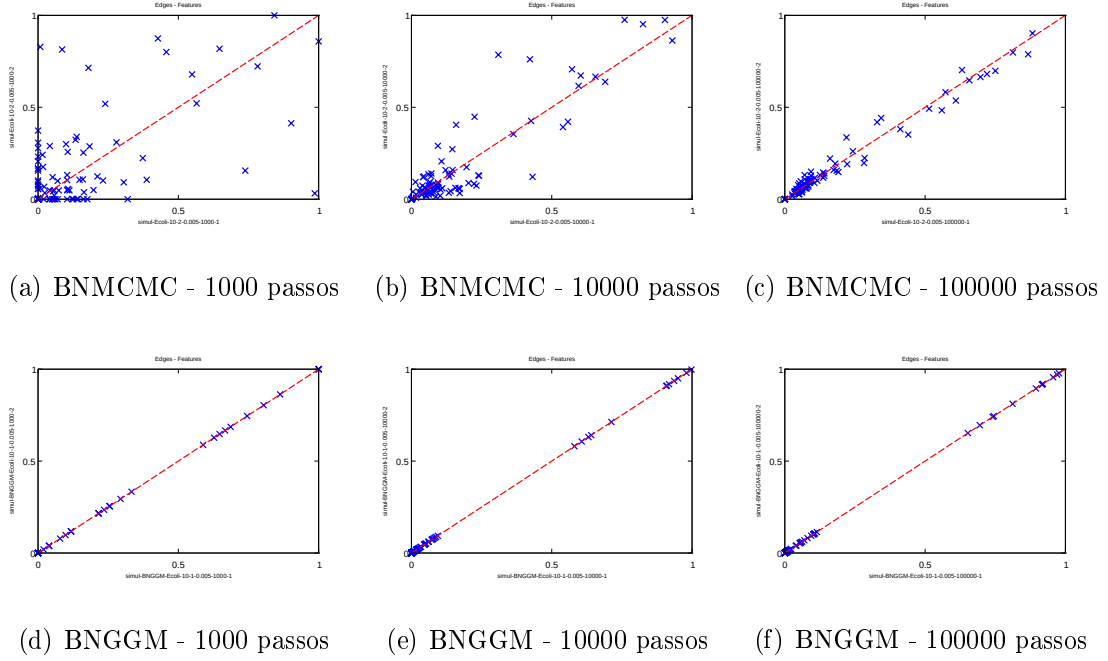


Figura 3.3: Estes gráficos representam o método visual heurístico, através do qual se avalia visualmente a convergência das simulações. Este método não é muito preciso devido ao fato de a análise ser feita visualmente e não numericamente. O painel superior da figura mostra os resultados para o método BNMCMC eo painel inferior mostra os resultados para o BNGGM. As subfiguras em cada painel, da esquerda para a direita, mostram respectivamente resultados para simulações com tamanho 1000, 10000, 100000. Neste método visual, quanto mais próximos os pontos estiverem em relação a reta, dizemos que a simulação convergiu. Exemplificando, na figura 3.3(a) pode-se se dizer que a simulação não convergiu, pois os pontos estão dispersos em relação a reta, já na figura 3.3(d) pode-se dizer que a simulação convergiu, pois os pontos encontram-se sobre a reta

como temos a linha definida por $y = x$, a equação 3.1 torna-se:

$$d_i = \frac{|x_i - y_i|}{\sqrt{2}} \quad (3.2)$$

Estamos interessados na distribuição dos pontos ao redor da linha $y = x$ e usamos a distância RMS como uma medida dessa distribuição. A distância RMS é definida por:

$$\text{rms} = \sqrt{\frac{1}{N} \sum_i^N (d_i)^2} \quad (3.3)$$

onde N é o número total de pontos e d_i é a distância do ponto i -th até a linha. Se combinarmos a equação 3.2 com 3.3 obtemos:

$$\text{rms} = \sqrt{\frac{1}{N} \sum_i^N \frac{|x_i - y_i|^2}{2}} \quad (3.4)$$

E então o valor obtido do cálculo da distância RMS (equação 3.4) é mostrado graficamente caracterizando o que chamamos no decorrer deste trabalho de gráfico da convergência. Um exemplo do gráfico que contém o cálculo da distância RMS é mostrado na figura 3.4.

3.3.3 Descrição Simulações

Para avaliar o novo método proposto foram utilizados três tipos de dados distintos que foram os dados gerados pela ferramenta GNW [Thomas Schaffter and Roulet, 2011], dados Gaussianos de uma distribuição Gaussiana Multivariável e dados reais gerados através de experimentos citométricos.

No total, temos à nossa disposição 15 conjuntos de dados. As simulações são obtidas a partir de cada um dos tipos de dados citados anteriormente, com cinco conjuntos de dados em cada tipo. Para cada um dos conjuntos de dados e para cada um dos métodos de inferência executamos 2 simulações. No total, realizando simulações com cada um dos métodos BNMCMC, BNGGM e BNMI para cada um dos 15 conjuntos, totalizando 90 simulações.

O número de passos das simulações foi ajustado para 10^4 a partir do qual foram coletadas amostras em intervalos de 10 passos.

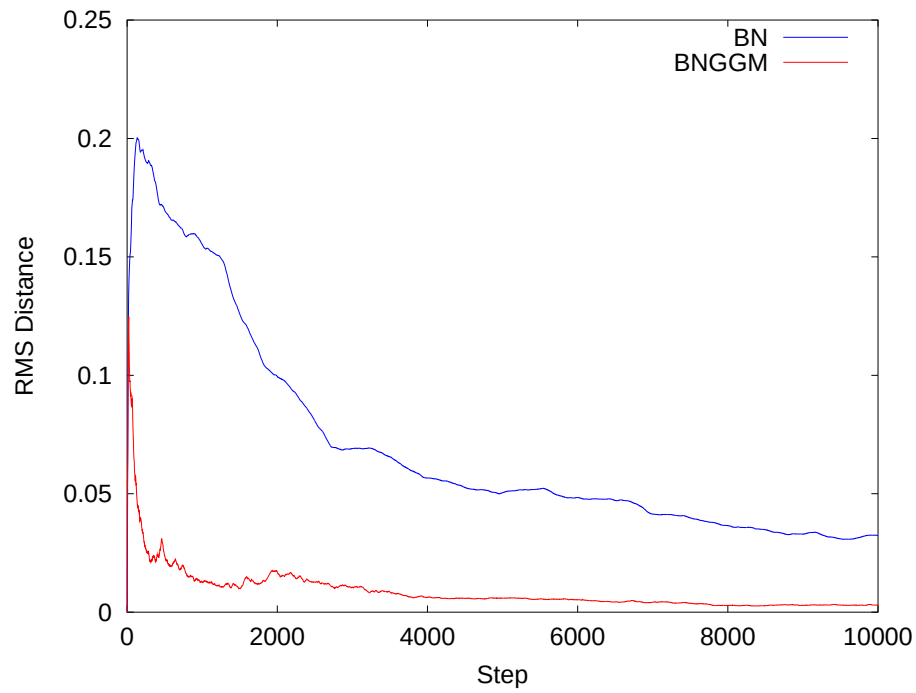


Figura 3.4: Este gráfico é baseado no método visual heurístico, porém é calculada a distância RMS que é uma medida da distribuição dos pontos ao redor da reta do gráfico visual.

Tipicamente, amostras iniciais do algoritmo de Metropolis-Hastings não são completamente válidas porque a Cadeia de Markov não se estabilizou, sendo assim, comumente a primeira metade das amostras da simulação MCMC é descartada na chamada fase de *burn-in*.

Capítulo 4

Resultados e Discussão

4.1 Método BNGGM

Nesta sessão são apresentados os resultados das simulações realizadas com os métodos BNMCMC e BNGGM. São utilizados os dados Ecoli, Gaussianos e Reais nestes experimentos. Todos os gráficos aqui apresentados são decorrentes dos resultados mostrados nos gráficos de curvas ROC que mostram as AUC obtidas através dos experimentos. Estes gráficos de curvas ROC encontram-se no apêndice A.

Os gráficos mostrados na figura 4.1 representam um resumo dos resultados das simulações. Nestes gráficos podemos observar a média e o desvio padrão dos valores das áreas resultantes das simulações realizadas com os métodos BNMCMC e BNGGM. Nas figura 4.1(a) e 4.1(b) podemos visualizar a média e desvio padrão das simulações realizadas com os dados Ecoli. As figuras 4.1(c) e 4.1(d) mostram a média e desvio padrão das simulações realizadas com os dados Gaussianos. Nas figuras 4.1(e) e 4.1(f) encontramos o resumo dos resultados das simulações realizadas com os dados Reais obtidos através dos experimentos de citometria de fluxo.

A figura 4.2 mostra os gráficos contendo o resumo da comparação entre os resultados das simulações realizadas com dados Ecoli através dos métodos BNMCMC e BNGGM. Cada ponto na curva mostrada nos gráficos das figuras 4.2(a), 4.2(b), 4.2(c), 4.2(d) e 4.2(e) representa a área resultante da simulação em cada um dos 10000 passos gerados.

A figura 4.3 mostra os gráficos contendo o resumo da comparação entre os resultados das simulações realizadas com dados Gaussianos através dos métodos BNMCMC e

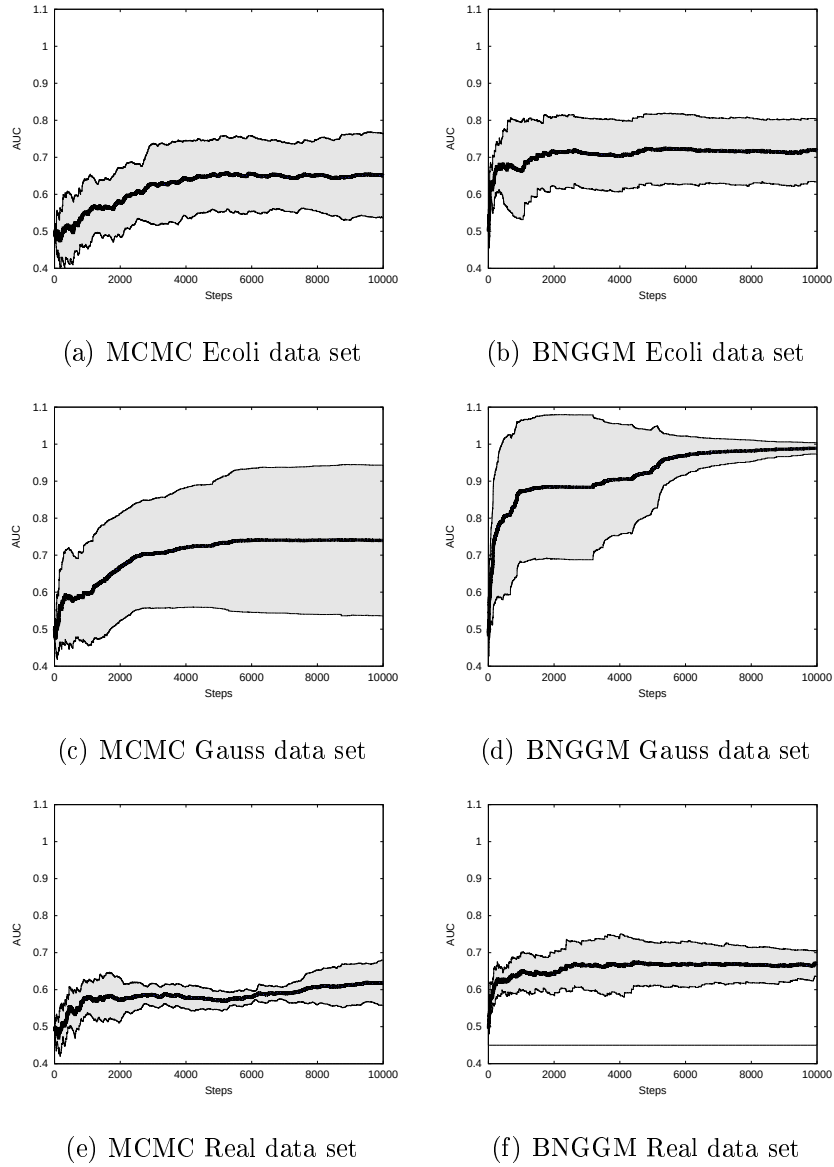


Figura 4.1: Estes gráficos apresentam a média das áreas resultantes das simulações, bem como o desvio padrão das mesmas: Na figura 4.1(a) são mostrados os gráficos da média e desvio padrão das áreas resultantes das simulações realizadas com o método MCMC e os dados Ecoli gerados pela ferramenta GNW; Na figura 4.1(b) apresenta-se o gráfico da média e desvio padrão resultante das simulações realizadas com o método BNGGM e os dados Ecoli gerados pela ferramenta GNW; Na figura 4.1(c) temos os gráficos da média e desvio padrão resultantes das simulações realizadas com o método MCMC e os dados Gaussianos; Na figura 4.1(d) temos o gráfico da média das áreas e desvio padrão resultante das simulações realizadas como método BNGGM e os dados Gaussianos; Na figura 4.1(e) apresenta-se o gráfico da média e desvio padrão resultante das simulações realizadas com o método MCMC e os dados Reais ;E na figura 4.1(f) temos o gráfico da média e desvio padrão das áreas resultantes das simulações realizadas com o método BNGGM e dados Reais

BNGGM. Cada ponto na curva mostrada nos gráficos das figuras 4.3(a), 4.3(b), 4.3(c), 4.3(d) e 4.3(e) representa a área resultante da simulação em cada um dos 10000 passos gerados.

Na figura 4.4 encontramos a comparação resumidamente entre os resultados das simulações de dados Reais dos métodos BNMCMC e BNGGM. Em cada ponto da curva encontrada nos gráficos das figuras 4.4(a), 4.4(b), 4.4(c), 4.4(d), 4.4(e) temos o valor da área para cada um dos 10000 passos.

Estes gráficos apresentados na figura 4.1 foram gerados baseados nos resultados contidos nas figuras 4.2, 4.3, 4.4, pois possuem os valores das áreas a cada passo de cada um dos 10000 passos de cada uma das simulações realizadas com os 5 conjuntos de dados dos tipos de dados Ecoli, Gaussiano e Reais, para o método novo utilizando o GGM como conhecimento a priori

Os gráficos contidos na figura 4.5 apresentam a convergência das simulações realizadas com os métodos BNMCMC e BNGGM. As figuras 4.5(a), 4.5(b) e 4.5(c) contém a convergência do experimento realizado com o conjunto de dados 1 de cada um dos tipos de dados Ecoli, Gaussianos e Reais.

As figuras 4.5(d), 4.5(e) e 4.5(f) contém a convergência do experimento realizado com o conjunto de dados 2 de cada um dos tipos de dados Ecoli, Gaussianos e Reais. Nas figuras 4.5(g), 4.5(h) e 4.5(i) temos a convergência do experimento realizado com o conjunto de dados 3 de cada um dos tipos de dados Ecoli, Gaussianos e Reais.

A convergência do experimento realizado com o conjunto de dados 4 de cada um dos tipos de dados Ecoli, Gaussianos e Reais encontra-se nas figuras 4.5(j), 4.5(k) e 4.5(l) e a convergência do experimento realizado com o conjunto de dados 5 de cada um dos tipos de dados Ecoli, Gaussianos e Reais encontra-se nas figuras 4.5(m), 4.5(n) e 4.5(o)

4.2 Método BNMI

Esta sessão apresenta os resultados das simulações realizadas com os métodos BNMCMC e BNMI. Aqui também são utilizados os dados Ecoli, Gaussianos e Reais nos experimentos, porém ao invés de utilizar os resultados das simulações realizadas com o GGM como conhecimento a priori, utilizamos os resultados da Informação Mútua como conhecimento

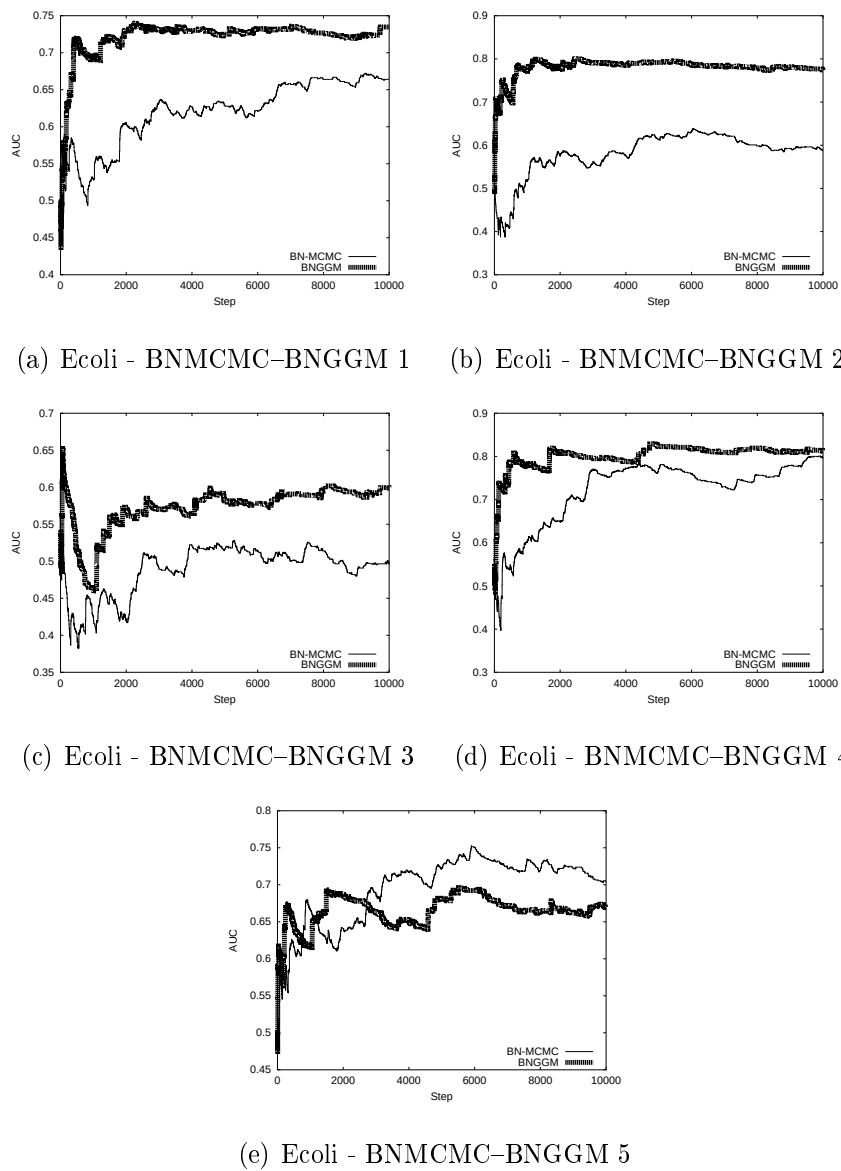
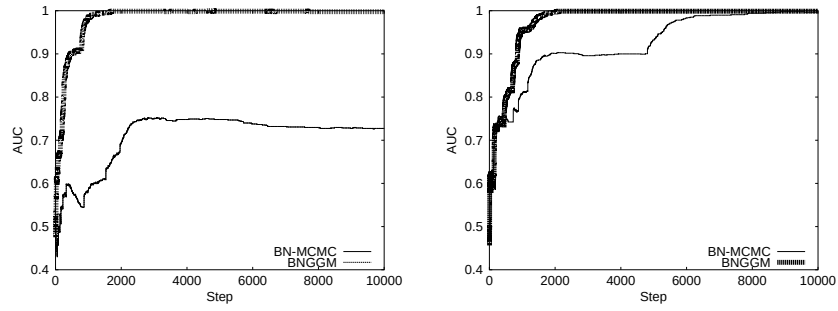
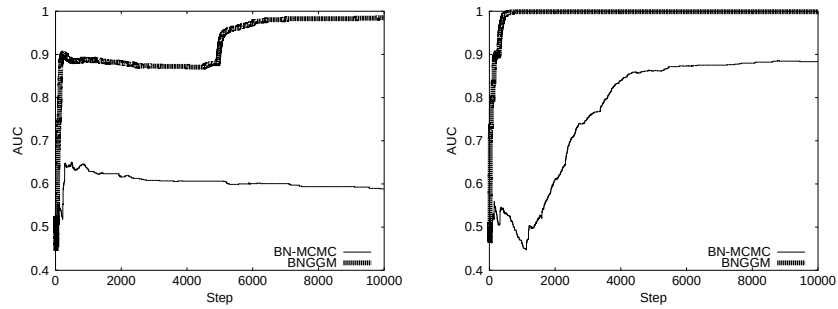


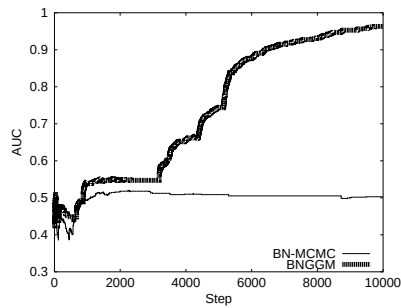
Figura 4.2: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BN-MCMC e BNGGM utilizando os dados Ecoli obtidos através da ferramenta GNW.



(a) Gauss - BN-MCMC-BNGGM 1 (b) Gauss - BN-MCMC-BNGGM 2

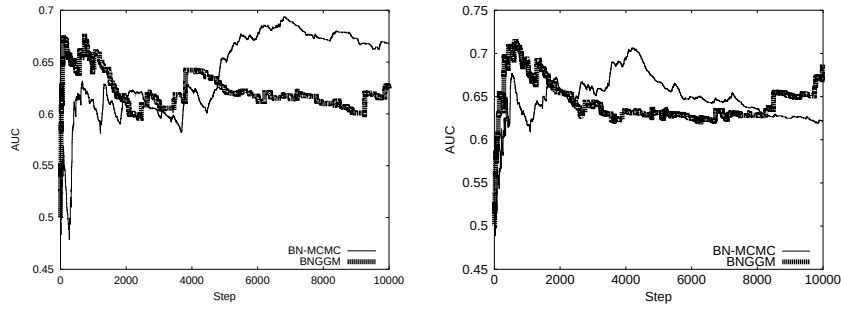


(c) Gauss - BN-MCMC-BNGGM 3 (d) Gauss - BN-MCMC-BNGGM 4

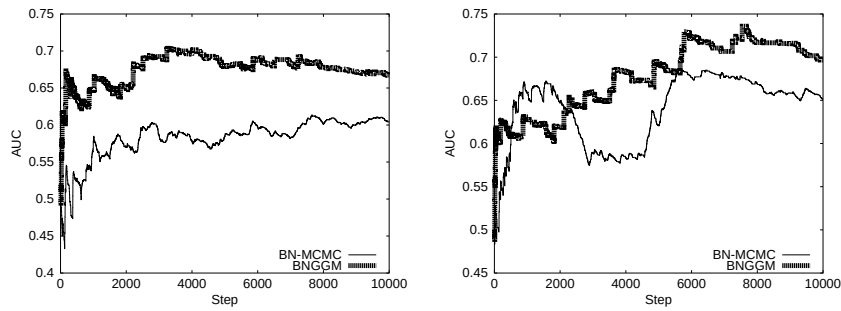


(e) Gauss - BN-MCMC-BNGGM 5

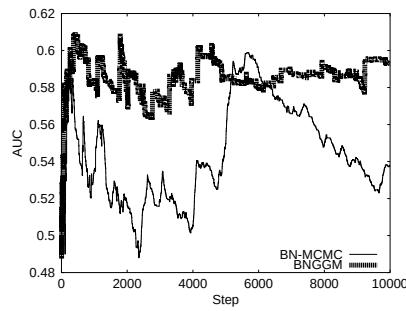
Figura 4.3: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BN-MCMC e BNGGM utilizando os dados Gaussianos.



(a) Reais - BN-MCMC-BNGGM 1 (b) Reais - BN-MCMC-BNGGM 2
teste



(c) Reais - BN-MCMC-BNGGM 3 (d) Reais - BN-MCMC-BNGGM 4



(e) Reais - BN-MCMC-BNGGM 5

Figura 4.4: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BN-MCMC e BNGGM utilizando os dados Reais.

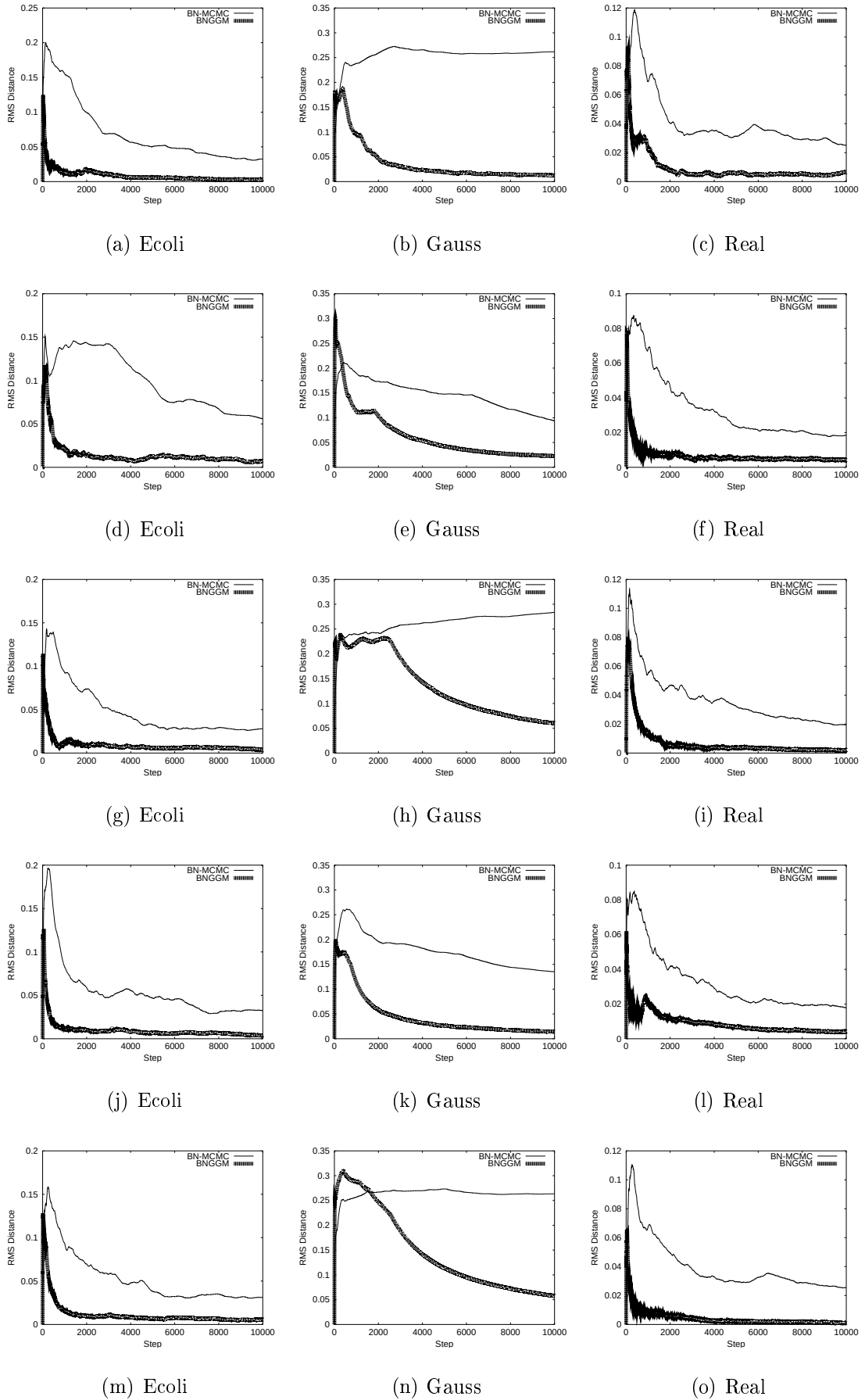


Figura 4.5: Estes gráficos mostram a comparação entre a convergência dos métodos BN-MCMC e BNGGM. As figuras contidas na primeira coluna representam a convergência das simulações realizadas com os 5 conjuntos de dados Ecoli, nas figuras da segunda coluna temos representada a convergência das simulações realizadas com os 5 conjuntos de dados Gaussianos e nas figuras da terceira coluna temos a convergência das simulações realizadas com 5 conjuntos de dados Reais

a priori.

Os gráficos contidos na figura 4.6 representam um resumo dos resultados das simulações realizadas com o método BNMCMC e com o método BNMI que utiliza o resultado das simulações realizadas com Informação Mútua como conhecimento a priori. Nas figuras 4.6(a) e 4.6(b) encontramos os resumos das simulações realizadas com os dados Ecoli. As figuras 4.6(c) e 4.6(d) mostram a média e desvio padrão das áreas obtidas através das simulações realizadas com os dados Gaussianos. E as figuras 4.6(e) e 4.6(f) mostram a média e desvio padrão das simulações realizadas com os dados Reais.

Estes gráficos apresentados na figura 4.6 foram gerados baseados nos resultados contidos nas figuras 4.7, 4.8 e 4.9, pois possuem os valores das áreas a cada passo de cada um dos 10000 passos de cada uma das simulações realizadas com os 5 conjuntos de dados dos tipos de dados Ecoli, Gaussiano e Reais, para o método novo utilizando a Informação Mútua como conhecimento a priori.

Na figura 4.7 temos os gráficos contendo o resumo da comparação entre os resultados das simulações realizadas com dados Ecoli através dos métodos BNMCMC e o método novo utilizando Informação Mútua BNMI. Cada ponto na curva mostrada nos gráficos das figuras 4.7(a), 4.7(b), 4.7(c), 4.7(d) e 4.7(e) representa a área resultante da simulação em cada um dos 10000 passos gerados.

A figura 4.8 mostra os gráficos contendo o resumo da comparação entre os resultados das simulações realizadas com dados Gaussianos através dos métodos BNMCMC e BNMI. Cada ponto na curva mostrada nos gráficos das figuras 4.8(a), 4.8(b), 4.8(c), 4.8(d) e 4.8(e) representa a área resultante da simulação em cada um dos 10000 passos gerados.

Encontramos a comparação resumidamente entre os resultados das simulações de dados Reais dos métodos BNMCMC e BNMI na figura 4.9. Em cada ponto da curva encontrada nos gráficos das figuras 4.9(a), 4.9(b), 4.9(c), 4.9(d), 4.9(e) temos o valor da área para cada um dos 10000 passos.

Os gráficos contidos na figura 4.10 apresentam a convergência das simulações realizadas com os métodos BNMCMC e BNMI, ou seja, utilizou-se Informação Mútua ao invés dos resultados do GGM. As figuras 4.10(a), 4.10(b) e 4.10(c) contêm a convergência do experimento realizado com o conjunto de dados 1 de cada um dos tipos de dados Ecoli, Gaussianos e Reais.

As figuras 4.10(d), 4.10(e) e 4.10(f) contém a convergência do experimento realizado com o conjunto de dados 2 de cada um dos tipos de dados Ecoli, Gaussianos e Reais. Nas figuras 4.10(g), 4.10(h) e 4.10(i) temos a convergência do experimento realizado com o conjunto de dados 3 de cada um dos tipos de dados Ecoli, Gaussianos e Reais.

A convergência do experimento realizado com o conjunto de dados 4 de cada um dos tipos de dados Ecoli, Gaussianos e Reais encontra-se nas figuras 4.10(j), 4.10(k) e 4.10(l), por último a convergência do experimento realizado com o conjunto de dados 5 de cada um dos tipos de dados Ecoli, Gaussianos e Reais encontra-se nas figuras 4.10(m), 4.10(n) e 4.10(o)

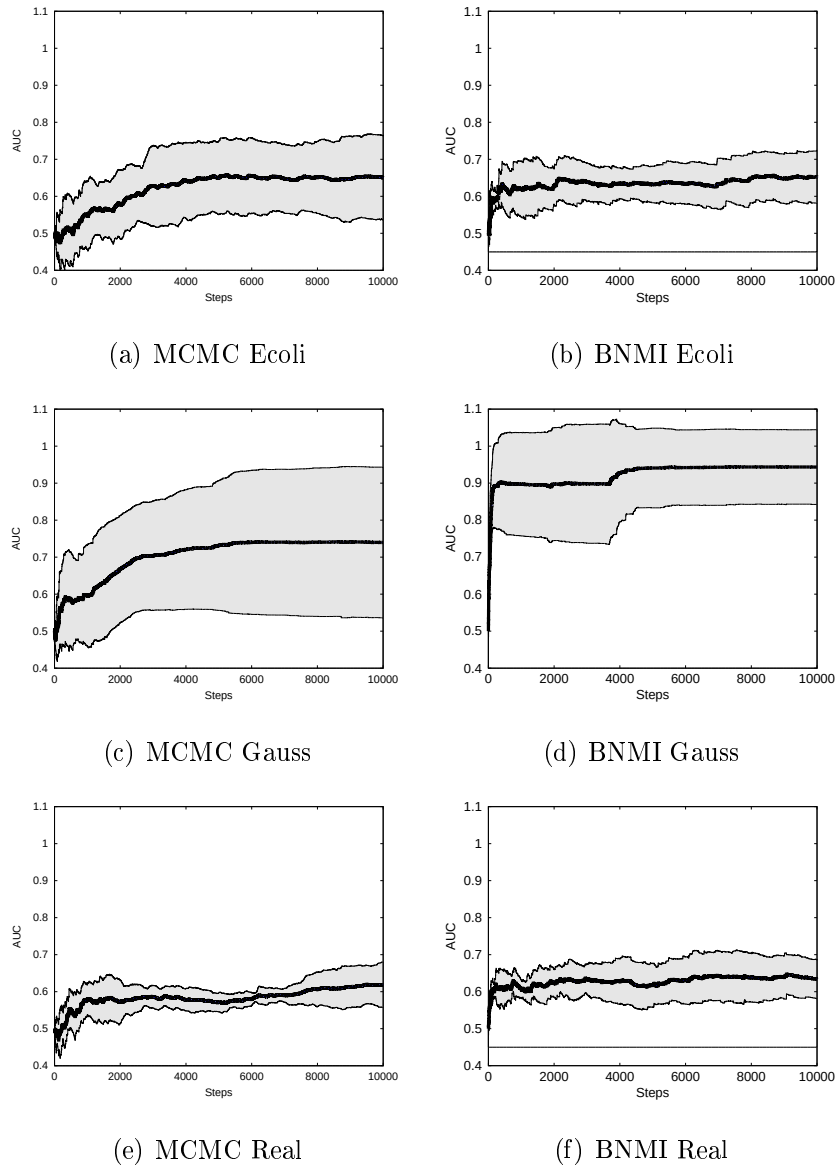


Figura 4.6: Estes gráficos apresentam a média das áreas resultantes das simulações, bem como o desvio padrão das mesmas: Na figura 4.6(a) são mostrados os gráficos da média e desvio padrão das áreas resultantes das simulações realizadas com o método MCMC e os dados Ecoli gerados pela ferramenta GNW; Na figura 4.6(b) apresenta-se o gráfico da média e desvio padrão resultante das simulações realizadas com o método BNMI e os dados Ecoli gerados pela ferramenta GNW; Na figura 4.6(c) temos os gráficos da média e desvio padrão resultantes das simulações realizadas com o método MCMC e os dados Gaussianos; Na figura 4.6(d) temos o gráfico da média das áreas e desvio padrão resultante das simulações realizadas como método BNMI e os dados Gaussianos; Na figura 4.6(e) apresenta-se o gráfico da média e desvio padrão resultante das simulações realizadas com o método MCMC e os dados Reais; E na figura 4.6(f) temos o gráfico da média e desvio padrão das áreas resultantes das simulações realizadas com o método BNMI e dados Reais

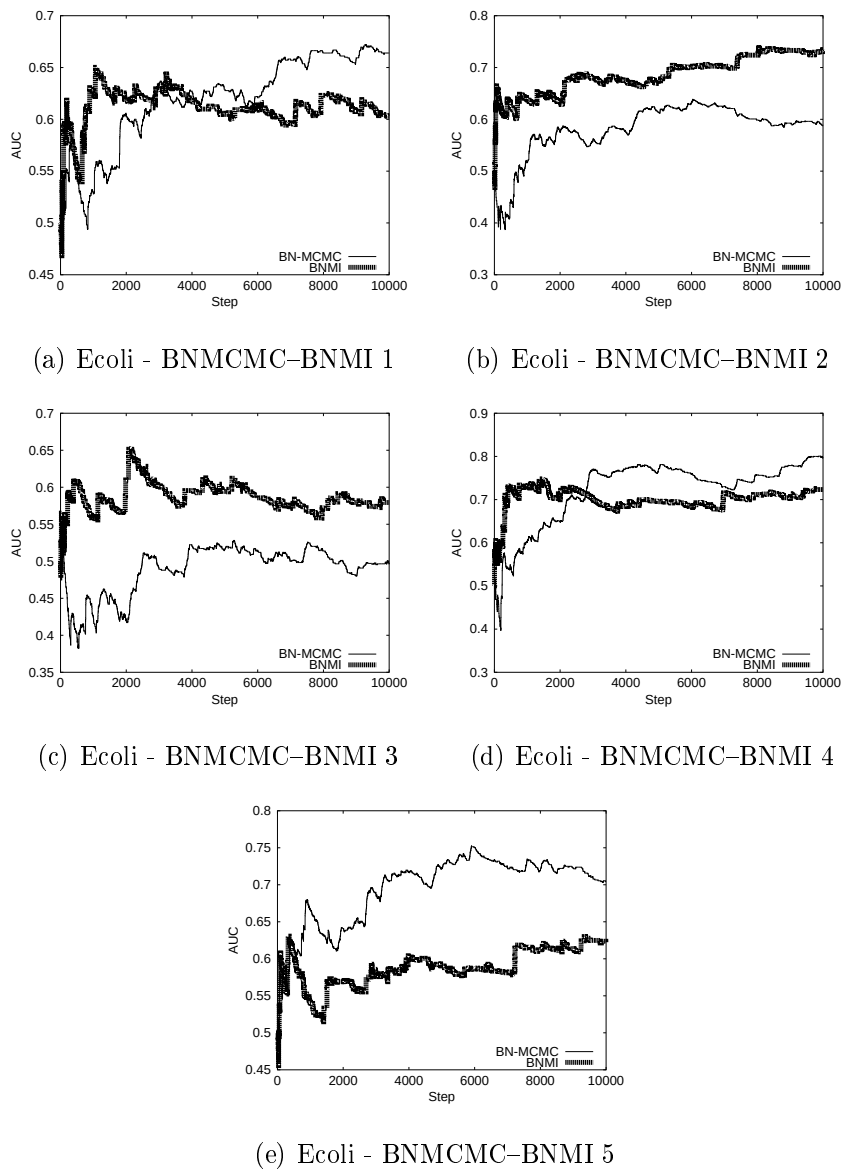
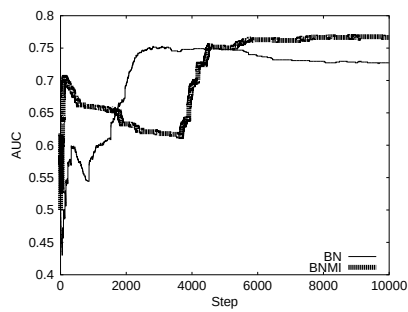
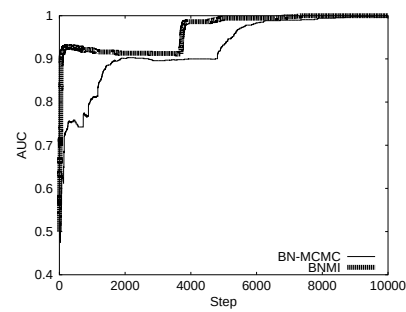


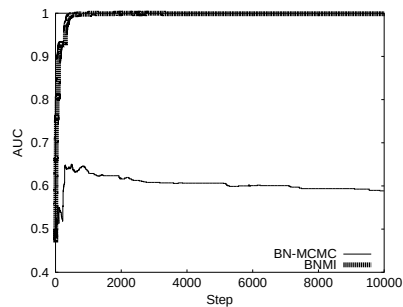
Figura 4.7: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BN-MCMC e BNMI utilizando os dados Ecoli obtidos através da ferramenta GNW.



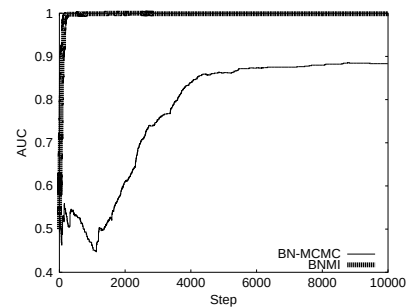
(a) Gauss - BN-MCMC-BNMI 1



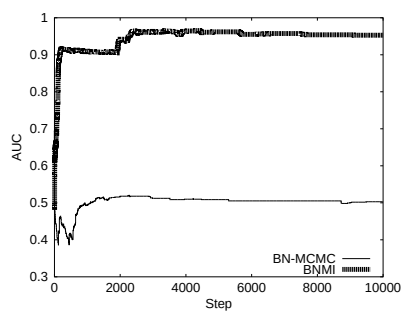
(b) Gauss - BN-MCMC-BNMI 2



(c) Gauss - BN-MCMC-BNMI 3

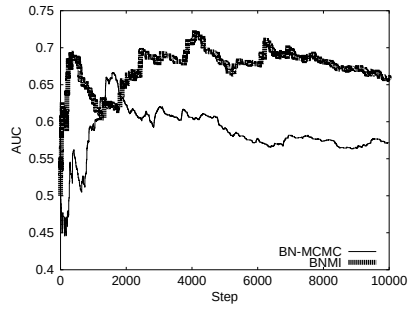


(d) Gauss - BN-MCMC-BNMI 4

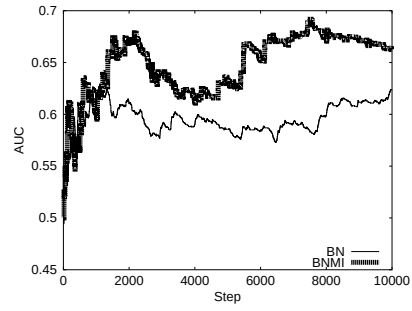


(e) Gauss - BN-MCMC-BNMI 5

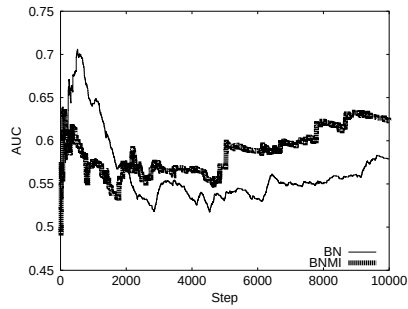
Figura 4.8: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BN-MCMC e BNMI utilizando os dados Gaussianos.



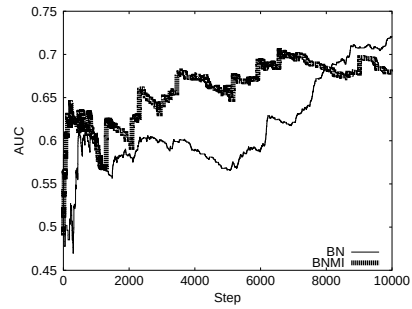
(a) Reais - BNMCMC-BNMI 1



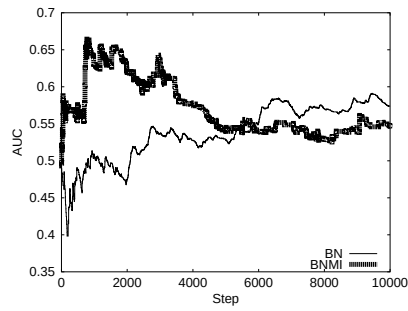
(b) Reais - BNMCMC-BNMI 2



(c) Reais - BNMCMC-BNMI 3



(d) Reais - BNMCMC-BNMI 4



(e) Reais - BNMCMC-BNMI 5

Figura 4.9: Estes gráficos apresentam as áreas resultantes das simulações realizadas com os métodos BNMCMC e BNMI utilizando os dados Reais.

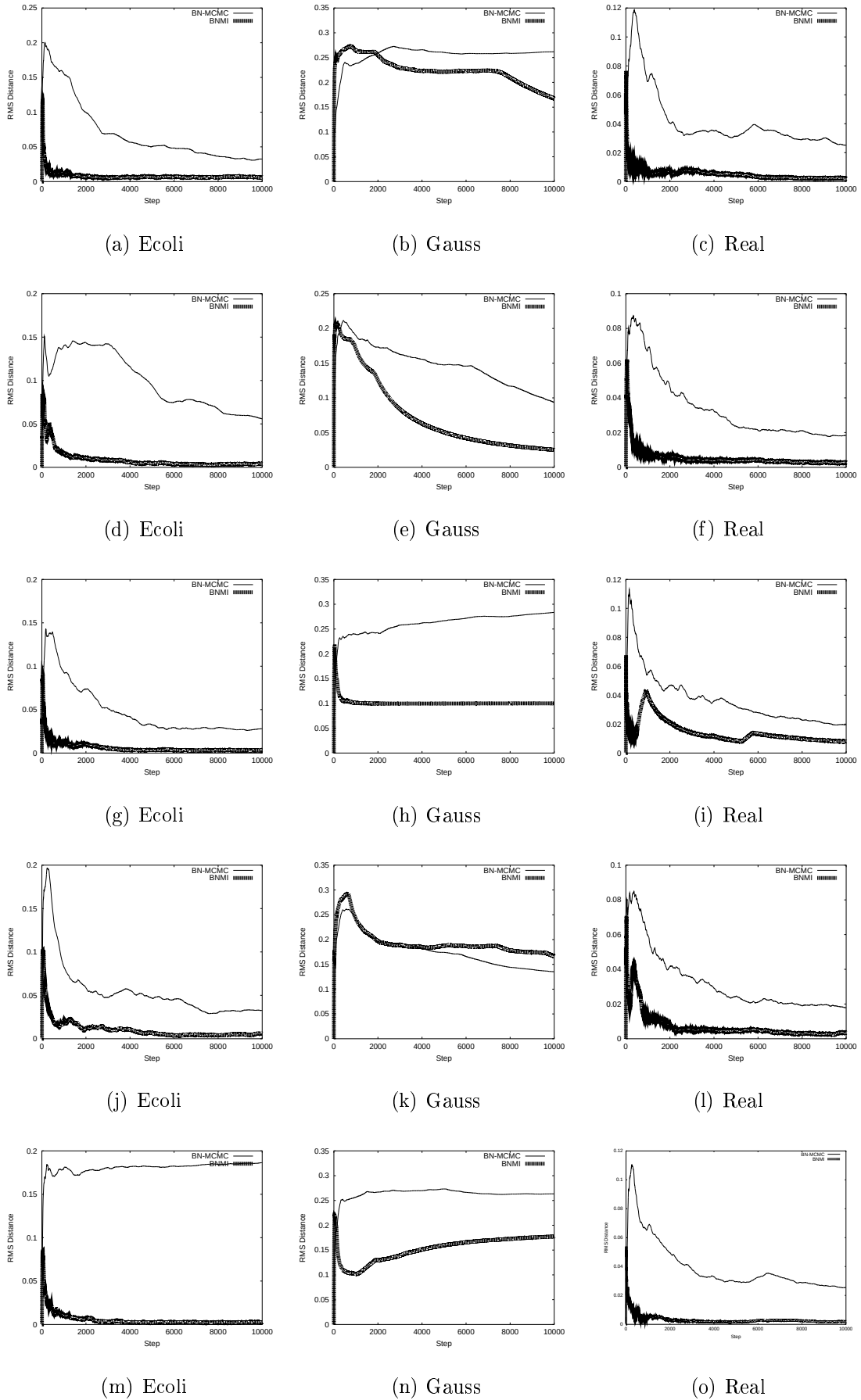


Figura 4.10: Estes gráficos mostram a comparação entre a convergência dos métodos MCMC e BNMI. As figuras contidas na primeira coluna representam a convergência das simulações realizadas com os 5 conjuntos de dados Ecoli, nas figuras da segunda coluna temos a convergência das simulações realizadas com os 5 conjuntos de dados Gaussianos e as figuras da terceira coluna representam a convergência das simulações realizadas com 5 conjuntos de dados Reais

Capítulo 5

Conclusões e Discussão

Através dos gráficos de resumo das Áreas abaixo da ROC (AUCs) podemos observar o quanto varia a área de cada simulação e podemos comparar os resumos das áreas resultantes das simulações realizadas com o método BNMCMC versus o método proposto BNGGM e também podemos realizar a comparação entre os resultados gerados mediante os métodos BNMCMC versus BNMI.

Comparando os gráficos 4.1(a) e 4.1(b) podemos observar nestes gráficos que as simulações em que utilizamos os dados Ecoli realizadas com o método BNGGM apresentam média maior que as simulações realizadas com o método BNMCMC e o desvio padrão das simulações realizadas com o método BNMCMC e com o método BNGGM é semelhante, assim podemos concluir que o método BNGGM utilizando dados Ecoli gerou AUCs com valores maiores, ou seja representa um preditor melhor que o método BNMCMC.

A comparação dos gráficos 4.1(c) e 4.1(d) mostra que as simulações em que utilizamos os dados Gaussianos realizadas com o método BNMCMC apresentam uma média das áreas AUCs muito menor que a média das áreas obtidas através do método BNGGM. Podemos observar também que o desvio padrão das simulações realizadas com o método BNMCMC se mantém equivalente ao desvio padrão das simulações realizadas com o método BNGGM até próximo do passo 6000, após podemos notar uma grande redução no desvio padrão do método BNGGM, assim podemos concluir que o método BNGGM continua, também neste caso, mostrando-se melhor do que o preditor do BNMCMC.

Ao comparar os gráficos 4.1(e) e 4.1(f) podemos observar que as simulações realizadas com o método BNMCMC apresentaram média menor do que as simulações realizadas

com o método BNGGM, mas também podemos observar que aproximadamente entre os passos 3000 e 7000 o método BNMCMC também resultou em um desvio padrão menor do que o do método BNGGM, após o passo 7000 podemos visualizar um aumento do desvio padrão do método BNMCMC e uma redução do desvio padrão do método BNGGM. Com os resultados desta comparação podemos então perceber que utilizando dados Reais o método BNGGM não mostrou-se melhor que o método o BNMCMC, porém este resultado era esperado devido aos ruídos contidos nestes tipos de dados.

Realizamos a análise do gráfico 4.6(a) em comparação com o gráfico 4.6(b) e observamos que nas simulações realizadas com os dados Ecoli as médias obtida através dos métodos BNMCMC e BNMI são equivalentes, porém também podemos visualizar que o desvio padrão do método BNMCMC apresenta-se maior que o desvio padrão do método BNMI, assim percebemos que o método BNMI também mostra ser um preditor melhor do que o método BNMCMC. E ainda ao verificar a média e desvio padrão do método BNGGM em comparação com a média e desvio e padrão do método BNMI percebemos que o método BNGGM mostra-se melhor do que o método BNMI mediante a avaliação das AUCs obtidas através da realização das simulações em que são utilizados os dados Ecoli.

Mediante a análise dos gráficos 4.1(c) e 4.6(d) podemos ver que média das AUCs do método BNMI utilizando dados Gaussianos mostra-se muito maior do que a média das AUCs do método BNMCMC ao utilizar os mesmos dados. Ao observar o desvio padrão de ambos os métodos vemos que o BNMCMC apresenta um desvio padrão muito maior do que o BNMI, sendo assim, o método proposto ao utilizar Informação Mútua como conhecimento a priori também mostra-se melhor do que o método BNMCMC.

Ao fazer a análise dos gráficos 4.6(e) e 4.6(f) visualizamos que os métodos BNMCMC e BNMI possuem a média das AUCs equivalente. Entre os passos aproximadamente 3000 e 7000 o BNMCMC possuem desvio padrão menor que o BNMI, mas ao final da simulação o desvio padrão do método BNMI mostra-se menor do que o BNMCMC. Podemos observar então que ao realizar as simulações com os dados Reais as AUCs produzidas são equivalentes em ambos os métodos que utilizam GGM e Informação Mútua respectivamente. Este resultado já era esperado devido a natureza complexa dos dados Reais que possuem muito ruído.

Nos gráficos contidos nas figuras 4.2, 4.3 e 4.4 são apresentados os valores das AUCs resultantes das simulações passo à passo. Ao analisarmos o gráfico 4.2 vemos que nas simulações realizadas com 4 dos conjuntos de dados Ecoli o método BNGGM sempre mostra-se com valores de AUC muito mais elevados que o método BNMCMC.

Na análise da figura 4.3 verificamos que em todas as simulações realizadas com os 5 conjuntos de dados Gaussianos o BNGGM apresenta valores de AUC muito maiores que os valores do BNMCMC. E finalmente ao observar a figura 4.4 também percebemos que nas simulações de 4 dos conjuntos de dados Ecoli o BNGGM gera AUCs com valores maiores do que o BNMCMC.

Realizamos também a análise dos gráficos que mostram passo à passo os resultados das simulações realizadas com o BNMI versus o BNMCMC. Aqui percebemos que dos 5 gráficos que encontram-se na figura 4.7 3 mostram que o BNMI gera AUCs maiores do que o BNMCMC. Na observação da figura 4.3 podemos visualizar através dos gráficos contidos nesta figura que o método BNMI ao utilizar dados Gaussianos gera AUCs maiores que o BNMCMC. Já ao analisar a figura 4.9 verificamos que o método BNMI gera AUCs maiores do que o BNMCMC nas simulações realizadas com 3 dos 5 conjuntos de dados Reais. Então, podemos concluir que as simulações realizadas com dados Reais em termos de análise de valores de AUCs geraram resultados piores do que as simulações realizadas com dados Ecoli e Gaussianos, independentemente de serem realizadas com BNMCMC, BNGGM ou BNMI.

Realizamos também análises referentes a convergência dos métodos BNMCMC, BNGGM e BNMI. Estas análises mostram os resultados visando o objetivo deste trabalho, que é o melhoramento da convergência.

Na figura 4.5 encontramos os gráficos que contém a comparação entre a convergência do BNMCMC e do BNGGM. Na primeira coluna encontramos a comparação entre os resultados das convergências dos métodos BNMCMC e BNGGM utilizando dados Ecoli, na segunda coluna encontramos os resultados das simulações que utilizam dados Gaussianos e finalmente na terceira coluna encontramos os resultados que envolvem dados Reais. Ao analisar a primeira coluna podemos verificar que o método BNGGM converge mais rápido que o BNMCMC ao realizar simulações com cada um dos 5 conjuntos de dados Ecoli. Nos resultados destas inferências encontramos o resultado esperado inicialmente, pois sendo o

propósito deste trabalho melhorar a convergência do método BNMCMC. Com as simulações realizadas com dados Ecoli, então ao analisar a convergência do BNMCMC e BNGGM, vemos que o BNGGM convergiu primeiramente que o método BNMCMC.

Na segunda coluna da figura 4.5 encontramos os valores das convergências resultantes das simulações realizadas com dados Gaussianos tanto com o BNMCMC quanto com o BNGGM. Nestes gráficos podemos observar que nos testes realizados com esses dados Gaussianos o BNGGM convergiu mais antecipadamente ainda do que nos testes realizados com dados Ecoli, sendo assim podemos concluir que o nosso método proposto utilizando GGM como conhecimento a priori gerou os resultados esperados, pois a convergência é muito melhor do que a convergência do BNMCMC.

Na terceira coluna encontramos a convergência das simulações realizadas com dados Reais e ainda obtivemos os resultados esperados, pois, embora tenhamos ruídos nos dados Reais, as simulações realizadas com os dados BNGGM tiveram uma convergência melhorada em relação ao BNMCMC.

Realizamos também a análise da convergência dos testes realizados com o BNMI, assim podemos comparar os métodos BNGGM e BNMI e analisar se há diferenças significativas entre utilizar GGM ou Informação Mútua como conhecimento a priori direcionador do BNMCMC.

Na primeira coluna da figura 4.10 encontramos os resultados obtidos através dos testes baseados em dados Ecoli. Analisando esta coluna podemos verificar que o BNMI também apresenta convergência melhor do que o BNMCMC. Na segunda coluna temos os resultados envolvendo dados Gaussianos e o método BNMI analogamente ao BNGGM apresenta convergência melhor do que o BNMCMC.

Encontramos na segunda coluna os resultados obtidos mediante os testes realizados com dados Gaussianos. Aqui podemos visualizar que os resultados já não foram tão bons quanto os resultados obtidos através do BNGGM, já que nem todos os experimentos realizados com os 5 conjuntos de dados apresentaram convergência melhor do que BNMCMC.

Por fim, na terceira coluna temos a apresentação da convergência das simulações realizadas com dados Reais. Nestes gráficos podemos observar que o BNMI convergiu mais rápido do que o BNMCMC nos experimentos com os 5 conjuntos de dados Reais.

Neste trabalho apresentamos um modelo hierárquico Bayesiano que por combinar

GGMs com BNs melhora a convergência do algoritmo BNMCMC aplicado à inferência de redes regulatórias.

Se os métodos BNMCMC, BNGGM e BNMI rodarem infinitamente fornecerão o mesmo resultado referente à exatidão da reconstrução da rede, pois amostram a distribuição posterior da rede e garantem a convergência na simulação infinita. A principal vantagem do método proposto não está relacionada com a precisão da reconstrução da rede, mas está relacionada com o número de passos necessários para alcançar a convergência.

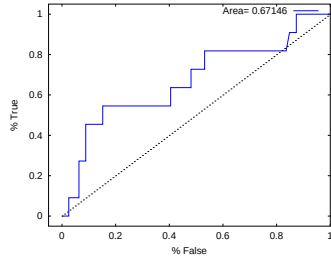
No decorrer do trabalho ao analisar os resultados obtidos podemos perceber que o objetivo deste trabalho foi alcançado, já que os gráficos de convergência nos mostram que o método proposto que utiliza o resultado das simulações do GGM como conhecimento a priori alcança a melhora da convergência, ou seja, mostra a convergência em muitos menos passos do que o método BNMCMC, assim como o método BNMI. Com isto, é possível executar simulações MCMC com menos iterações. É importante notar que a informação extra utilizada no BNGGM e no BNMI é obtida a partir dos dados em si, portanto, não há necessidade de qualquer outra fonte de informação.

Também percebemos que embora a convergência melhore, não obtivemos grandes melhoras nos valores das áreas AUC. Assim propomos como trabalho futuro o aperfeiçoamento do método proposto de maneira que possamos obter melhores resultados não somente na convergência do método.

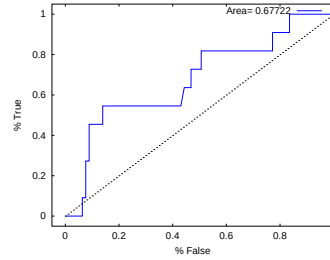
Apêndice A

Gráficos Curvas ROC

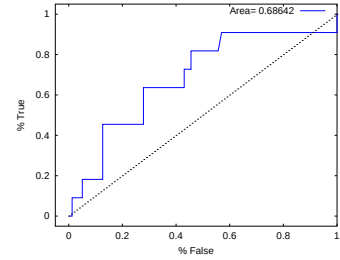
Neste apêndice apresentamos os gráficos AUC das 90 simulações realizadas. A figura A.1 mostra as áreas das simulações realizadas com dados Ecoli (GNW) e com método BN-MCMC, nas figuras A.2 e A.3 apresentamos as áreas obtidas através dos experimentos feitos com o método BN-MCMC com os dados Gaussianos e Reais respectivamente. Nas figuras A.4, A.5 e A.6 encontramos as áreas obtidas através das simulações realizadas com o método BNGGM utilizando os dados Ecoli (GNW), Gaussianos e Reais respectivamente. Por último, nas figuras A.7, A.8 e A.9 apresentamos as áreas obtidas pelos experimentos realizados com os mesmos dados descritos anteriormente, porém utilizamos o método BNMI. Os demais gráficos apresentados neste trabalho são decorrentes dos resultados apresentados neste apêndice.



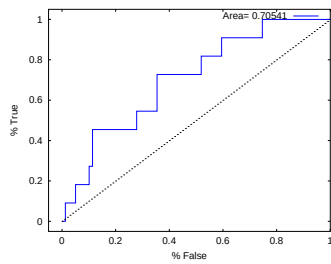
(a) MCMC - Ecoli 1 - 1



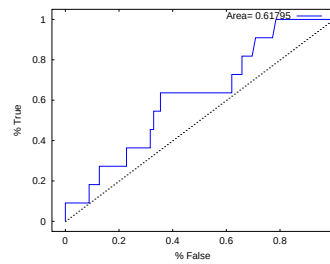
(b) MCMC - Ecoli 1 - 2



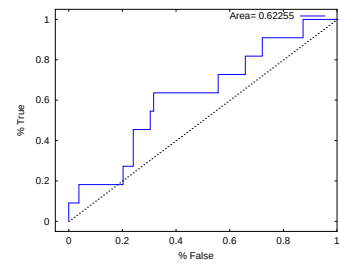
(c) MCMC - Ecoli 2 - 1



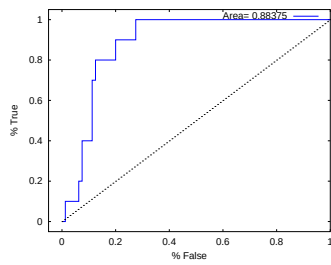
(d) MCMC - Ecoli 2 - 2



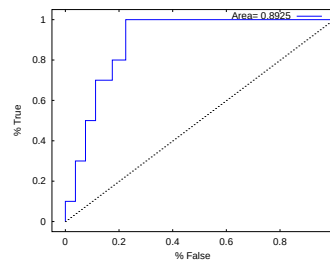
(e) MCMC - Ecoli 3 - 1



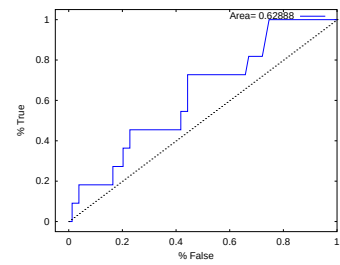
(f) MCMC - Ecoli 3 - 2



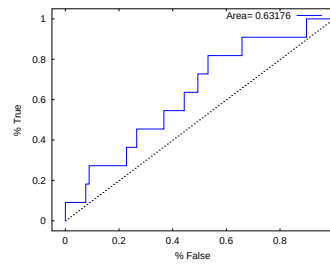
(g) MCMC - Ecoli 4 - 1



(h) MCMC - Ecoli 4 - 2

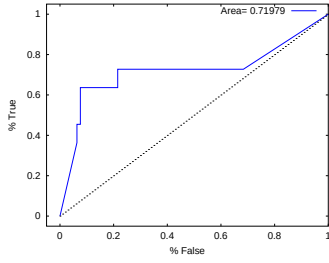


(i) MCMC - Ecoli 5 - 1

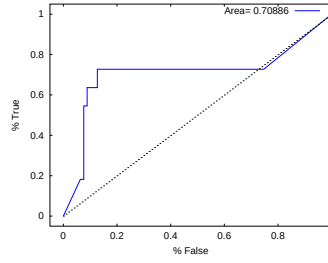


(j) MCMC - Ecoli 5 - 2

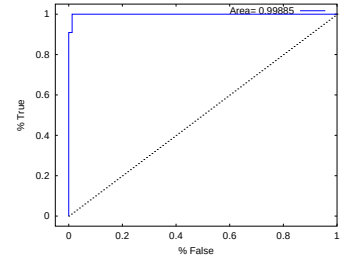
Figura A.1: Estes gráficos mostram as áreas resultantes das simulações com o método MCMC e os dados gerados pela ferramenta GNW.



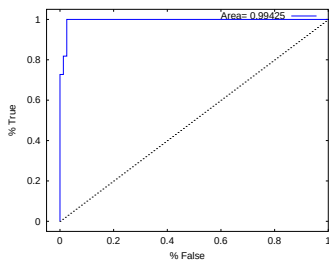
(a) MCMC - Gauss 1 - 1



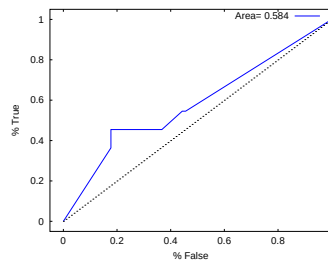
(b) MCMC - Gauss 1 - 2



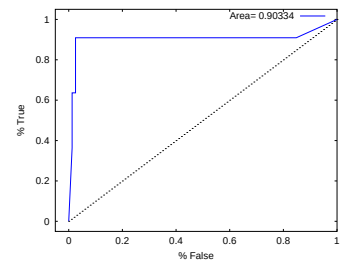
(c) MCMC - Gauss 2 - 1



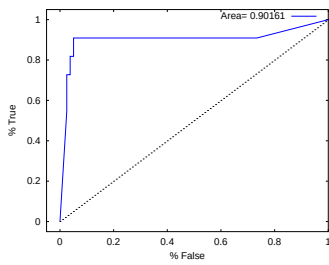
(d) MCMC - Gauss 2 - 2



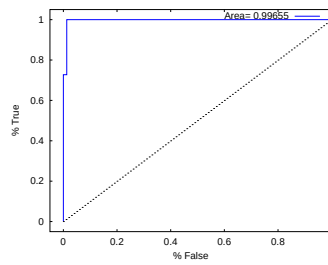
(e) MCMC - Gauss 3 - 1



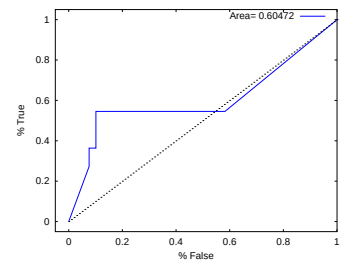
(f) MCMC - Gauss 3 - 2



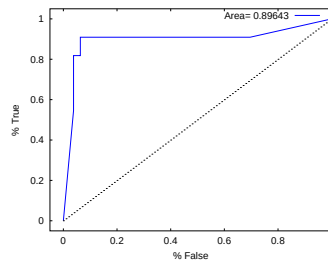
(g) MCMC - Gauss 4 - 1



(h) MCMC - Gauss 4 - 2

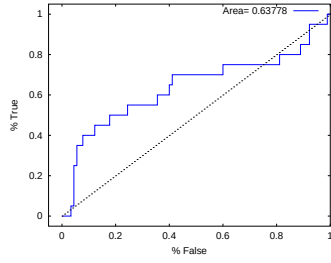


(i) MCMC - Gauss 5 - 1

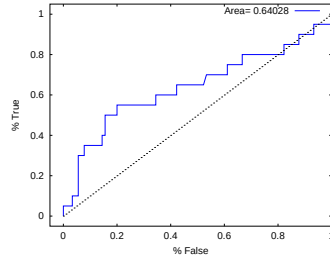


(j) MCMC - Gauss 5 - 2

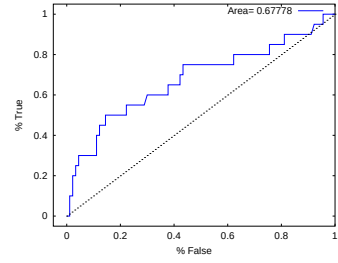
Figura A.2: Estes gráficos mostram as áreas resultantes das simulações com o método MCMC e os dados Gaussianos.



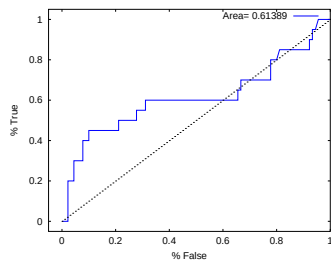
(a) MCMC - Real 1 - 1



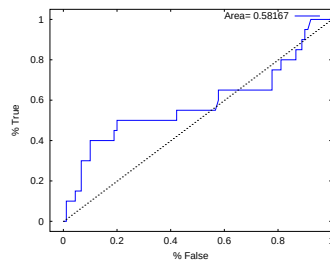
(b) MCMC - Real 1 - 2



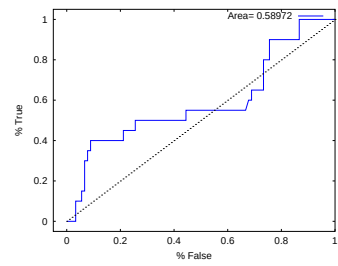
(c) MCMC - Real 2 - 1



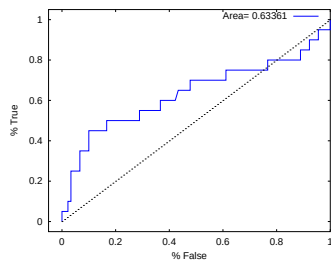
(d) MCMC - Real 2 - 2



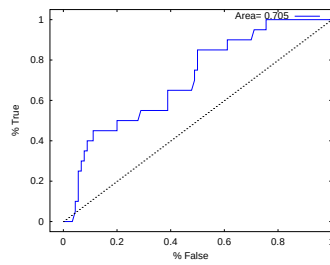
(e) MCMC - Real 3 - 1



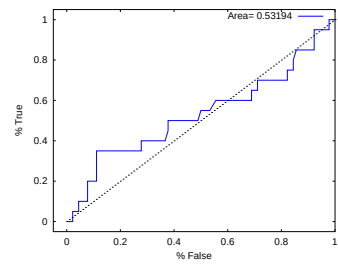
(f) MCMC - Real 3 - 2



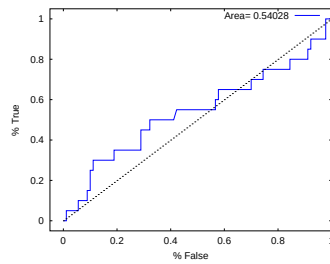
(g) MCMC - Real 4 - 1



(h) MCMC - Real 4 - 2

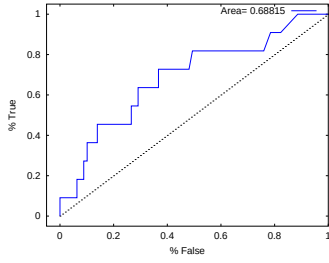


(i) MCMC - Real 5 - 1

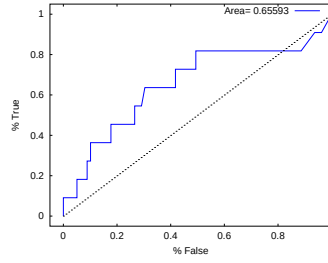


(j) MCMC - Real 5 - 2

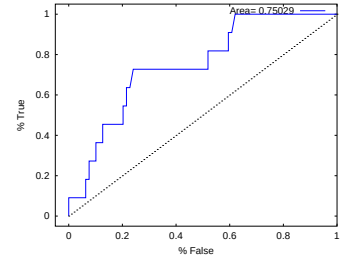
Figura A.3: Estes gráficos mostram as áreas resultantes das simulações com o método MCMC e os dados Reais.



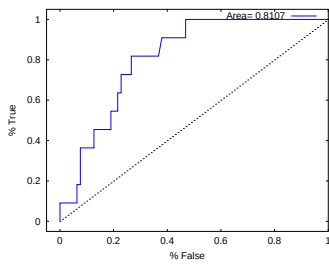
(a) BNGGM - Ecoli 1 - 1



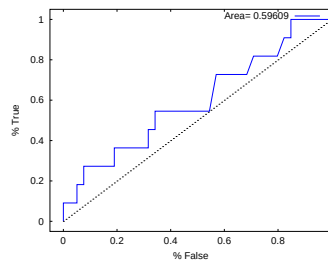
(b) BNGGM - Ecoli 1 - 2



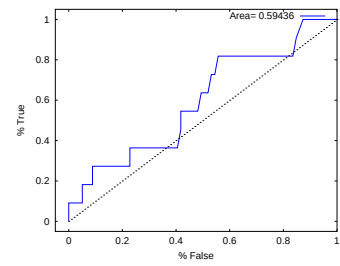
(c) BNGGM - Ecoli 2 - 1



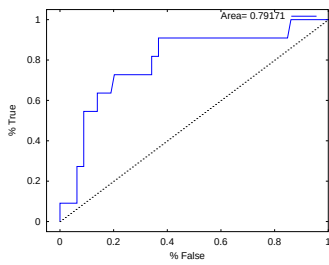
(d) BNGGM - Ecoli 2 - 2



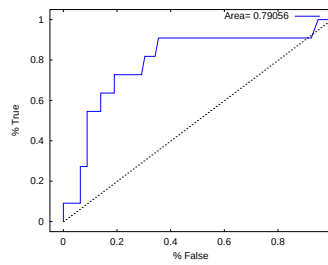
(e) BNGGM - Ecoli 3 - 1



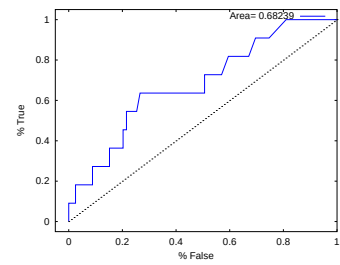
(f) BNGGM - Ecoli 3 - 2



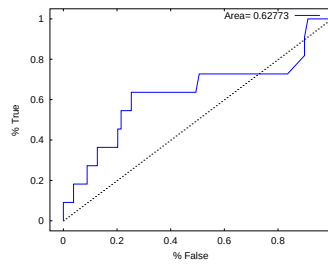
(g) BNGGM - Ecoli 4 - 1



(h) BNGGM - Ecoli 4 - 2

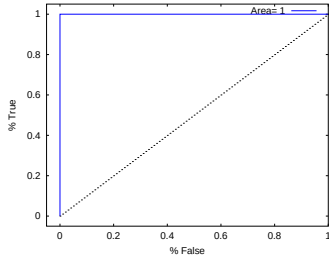


(i) BNGGM - Ecoli 5 - 1

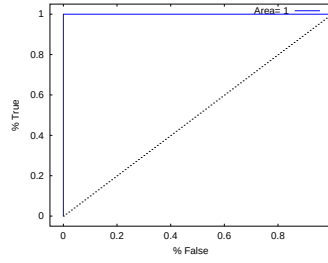


(j) BNGGM - Ecoli 5 - 2

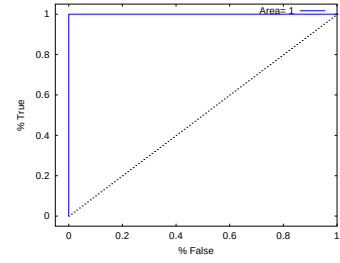
Figura A.4: Estes gráficos mostram as áreas resultantes das simulações com o método BNGGM e os dados gerados pela ferramenta GNW.



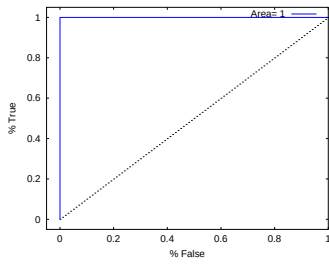
(a) BNGGM - Gauss 1 - 1



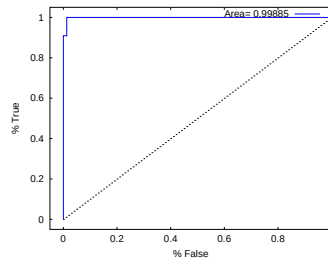
(b) BNGGM - Gauss 1 - 2



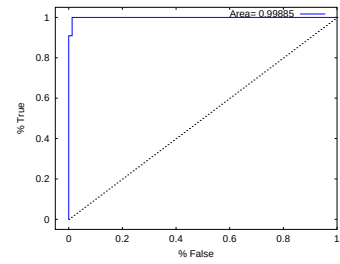
(c) BNGGM - Gauss 2 - 1



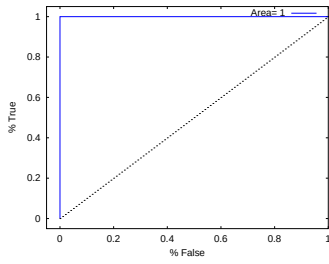
(d) BNGGM - Gauss 2 - 2



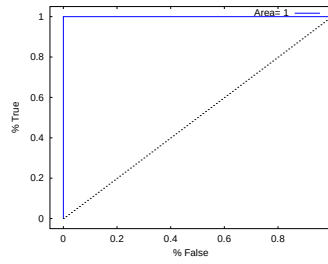
(e) BNGGM - Gauss 3 - 1



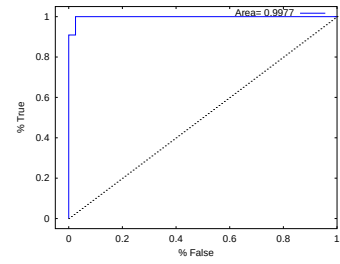
(f) BNGGM - Gauss 3 - 2



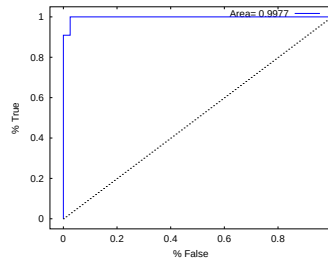
(g) BNGGM - Gauss 4 - 1



(h) BNGGM - Gauss 4 - 2

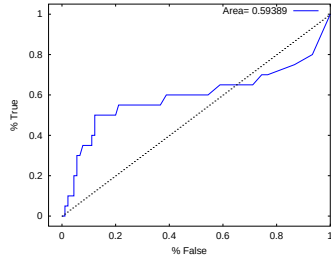


(i) BNGGM - Gauss 5 - 1

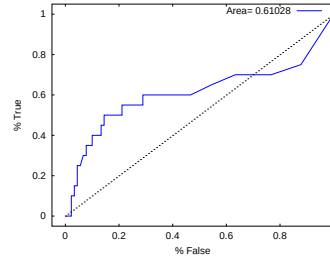


(j) BNGGM - Gauss 5 - 2

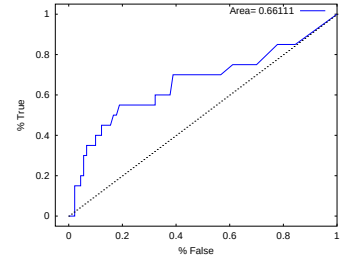
Figura A.5: Estes gráficos mostram as áreas resultantes das simulações com o método BNGGM e os dados Gaussianos.



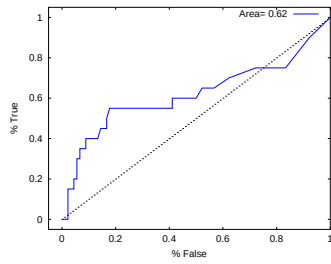
(a) BNGGM - Real 1 - 1



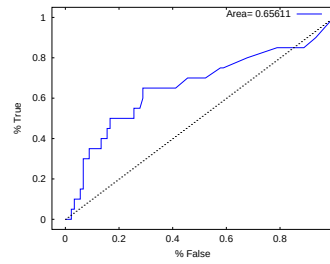
(b) BNGGM - Real 1 - 2



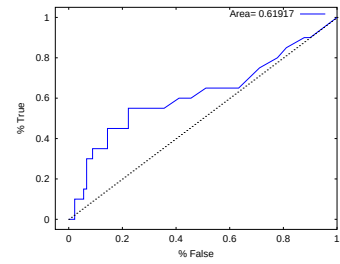
(c) BNGGM - Real 2 - 1



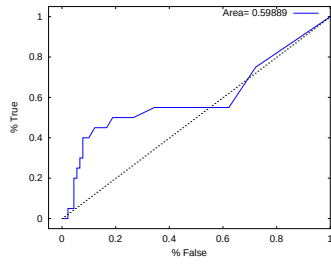
(d) BNGGM - Real 2 - 2



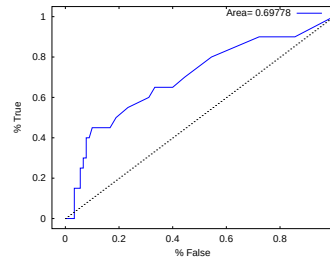
(e) BNGGM - Real 3 - 1



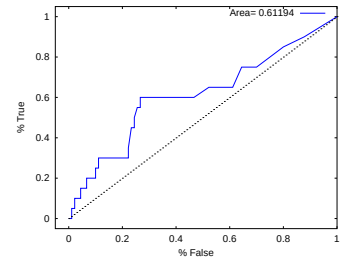
(f) BNGGM - Real 3 - 2



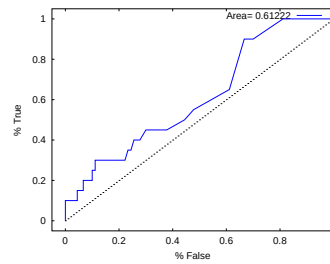
(g) BNGGM - Real 4 - 1



(h) BNGGM - Real 4 - 2

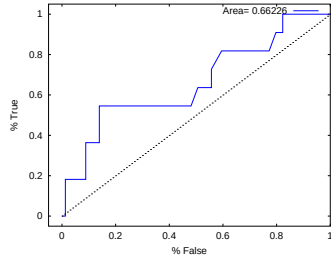


(i) BNGGM - Real 5 - 1

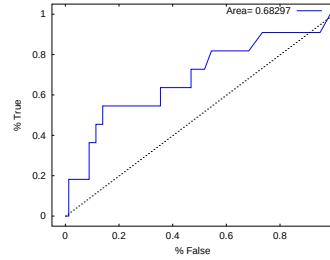


(j) BNGGM - Real 5 - 2

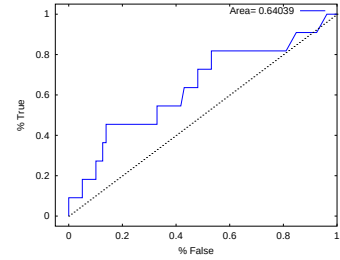
Figura A.6: Estes gráficos mostram as áreas resultantes das simulações com o método BNGGM e os dados Reais.



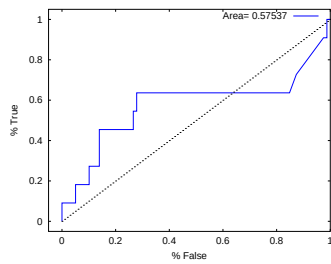
(a) BNMI - Ecoli 1 - 1



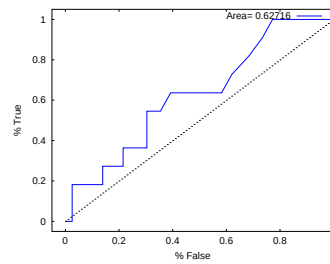
(b) BNMI - Ecoli 1 - 2



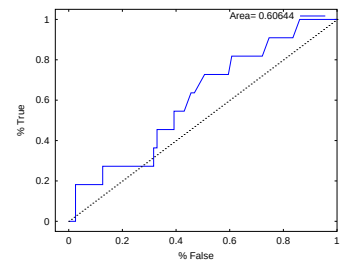
(c) BNMI - Ecoli 2 - 1



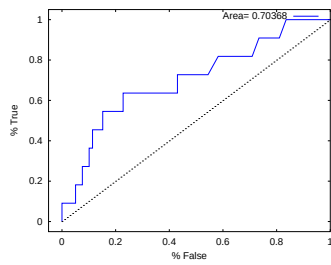
(d) BNMI - Ecoli 2 - 2



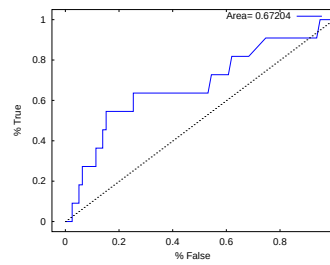
(e) BNMI - Ecoli 3 - 1



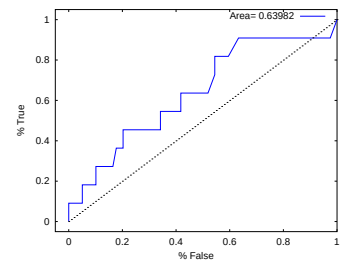
(f) BNMI - Ecoli 3 - 2



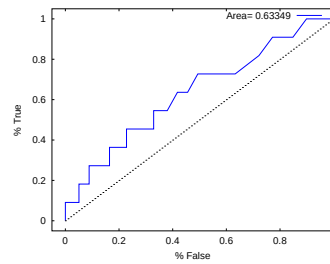
(g) BNMI - Ecoli 4 - 1



(h) BNMI - Ecoli 4 - 2

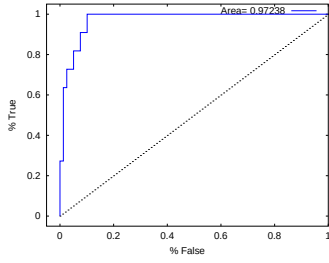


(i) BNMI - Ecoli 5 - 1

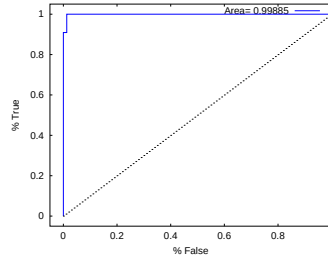


(j) BNMI - Ecoli 5 - 2

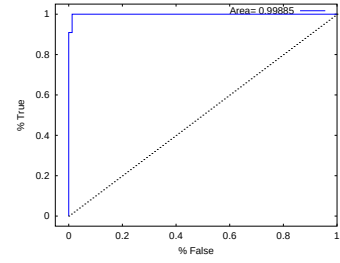
Figura A.7: Estes gráficos mostram as áreas resultantes das simulações com o método BNMI e os dados gerados pela ferramenta GNW.



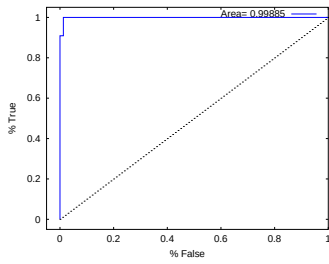
(a) BNMI - Gauss 1 - 1



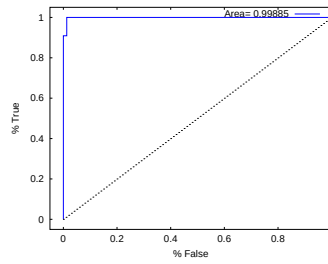
(b) BNMI - Gauss 1 - 2



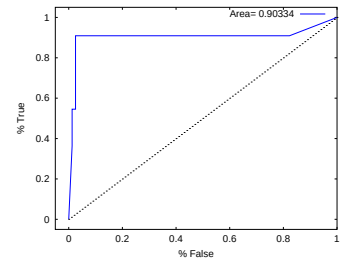
(c) BNMI - Gauss 2 - 1



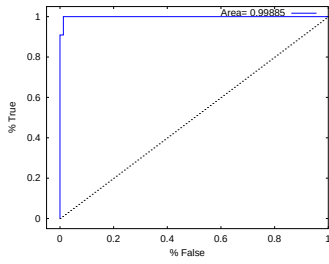
(d) BNMI - Gauss 2 - 2



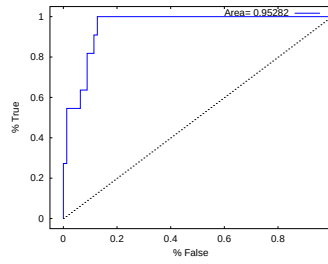
(e) BNMI - Gauss 3 - 1



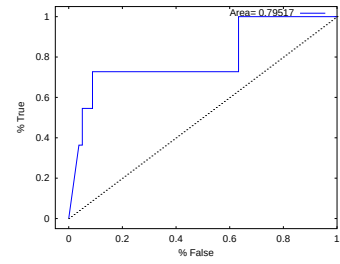
(f) BNMI - Gauss 3 - 2



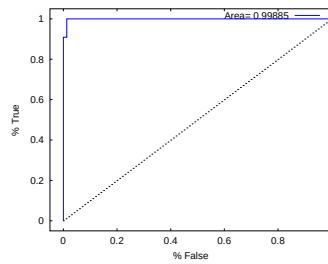
(g) BNMI - Gauss 4 - 1



(h) BNMI - Gauss 4 - 2

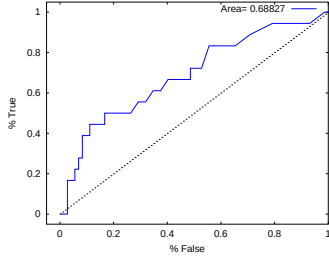


(i) BNMI - Gauss 5 - 1

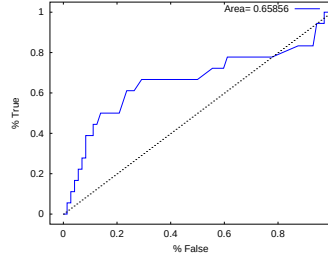


(j) BNMI - Gauss 5 - 2

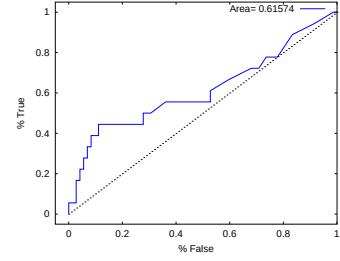
Figura A.8: Estes gráficos mostram as áreas resultantes das simulações com o método BNMI e os dados Gaussianos.



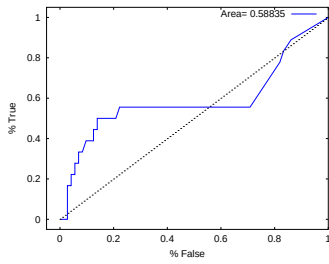
(a) BNMI - Real 1 - 1



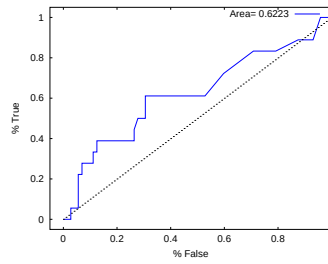
(b) BNMI - Real 1 - 2



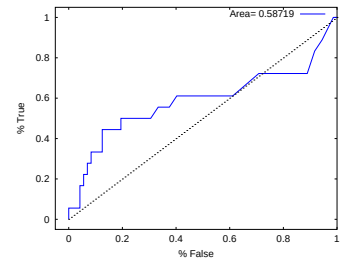
(c) BNMI - Real 2 - 1



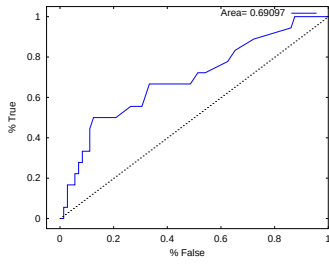
(d) BNMI - Real 2 - 2



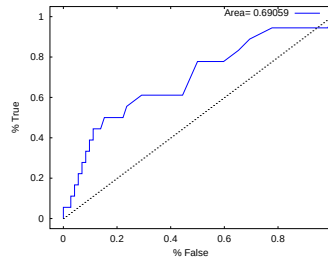
(e) BNMI - Real 3 - 1



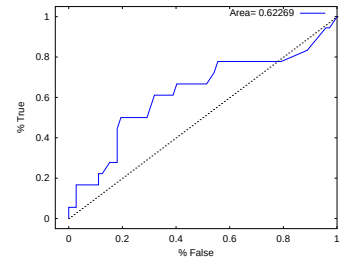
(f) BNMI - Real 3 - 2



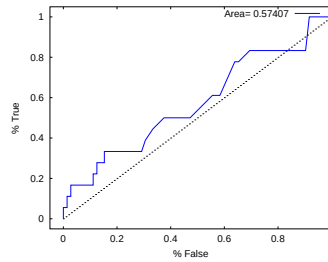
(g) BNMI - Real 4 - 1



(h) BNMI - Real 4 - 2



(i) BNMI - Real 5 - 1



(j) BNMI - Real 5 - 2

Figura A.9: Estes gráficos mostram as áreas resultantes das simulações com o método BNMI e os dados Reais.

Bibliografia

- [Adriano V. Werhli and Husmeier, 2006] Adriano V. Werhli, M. G. and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22:2523–2531.
- [Artuso, 2011] Artuso, A. R. (2011). Entropias de shannon e rényi aplicadas ao reconhecimento de padrões. *Revista CIATEC - UPF*.
- [Cowles and Carlin, 2006] Cowles, M. and Carlin, B. (2006). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904.
- [Edwards, 2000] Edwards, D. M. (2000). *Introduction to Graphical Modelling*. Springer Verlag, New York.
- [Friedman and Koller, 2001] Friedman, N. and Koller, D. (2001). Being bayesian about network structure - a bayesian approach to structure discovery in bayesian networks. *Kluwer Academic Publishers. Printed in the Netherlands*.
- [Friedman et al., 2000] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620.
- [Friedman et al., 1998] Friedman, N., Murphy, K., and Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In Cooper, G. F. and Moral, S., editors, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence(UAI)*, pages 139–147, San Francisco, CA. Morgan Kaufmann.

- [Gelman, 1996] Gelman, A. (1996). *Inference and Monitoring Convergence*, pages 131–143. Chapman & Hall, London.
- [Gibas and Jambeck, 2002] Gibas, C. and Jambeck, P. (2002). Developing bioinformatics computer skills. *Yale Journal of Biology and Medicine*.
- [Hastings, 1970] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.
- [Heckerman, 1994] Heckerman, D. (1994). Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, Redmond, Washington.
- [Heckerman, 1995] Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington,.
- [Ilya Shmulevich and Zhang, 2002] Ilya Shmulevich, Edward R. Dougherty, S. K. and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18.
- [Imoto et al., 2003] Imoto, S., Higuchi, T., Goto, T., Tashiro, Kousuke and Kuhara, S., and Miyano, S. (2003). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings IEEE Computer Society Bioinformatics Conference*, (CSB’03):104–113.
- [Ledoit and Wolf, 2004] Ledoit, O. and Wolf, M. (2004). A well conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.
- [Lopes, 2011] Lopes, F. M. (2011). Redes complexas de expressão gênica: síntese, identificação, análise e aplicações. Master’s thesis, Universidade de São Paulo.
- [Madigan and York, 1995] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232.
- [Paninski, 2003] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*.

- [Raftery and Lewis, 1995] Raftery, A. and Lewis, S. (1995). *The number of iterations, convergence diagnostics and generic Metropolis algorithms*, pages 115–130. Chapman & Hall, London.
- [Sachs et al., 2005] Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.
- [Schaefer and Strimmer, 2005] Schaefer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4. Article 32.
- [Shawn Martin et al., 2005] Shawn Martin, Z. Z., Martino, A., and Faulon, J.-L. (2005). Boolean dynamics of genetic regulatory networks inferred from microarray time series data.
- [Thomas Schaffter and Roulet, 2006] Thomas Schaffter, D. M. and Roulet, G. (2006). Dialogue for reverse engineering assessments and methods. <http://www.the-dream-project.org/>.
- [Thomas Schaffter and Roulet, 2011] Thomas Schaffter, D. M. and Roulet, G. (2011). Genenetweaver. <http://gnw.sourceforge.net/genenetweaver.html>.
- [Toni and Stumpf, 2010] Toni, T. and Stumpf, M. P. H. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26.
- [Werhli, 2007a] Werhli, A. V. (2007a). *Reconstruction of gene regulatory networks from postgenomic data*. PhD thesis, School of Informatics University of Edinburgh.
- [Werhli, 2007b] Werhli, A. V. (2007b). *Reconstruction of gene regulatory networks from postgenomic data*. PhD thesis, The University of Edinburgh, Edinburgh.
- [Werhli and Husmeier, 2007] Werhli, A. V. and Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6.