Scoring Functions
for Bayesian network structure learning

Hoai-Tuong NGUYEN
Polytechnic School of Nantes University
hoai-tuong.nguyen@etu.univ-nantes.fr

April 28, 2008

Abstract. In the construction of the Bayesian network (BN), the question of how to compare different structures is very important. This is the problem of selection a BN from a collection of observed data. This problem is known as the model selection problem. In the literature of model selection, there are many methods for decision a "true model". One of them is to maximize the scoring function according to the hypothesis that the observation data was generated and the probability distribution of this data. Inspired by that, *this paper presents a bibliography work about the scoring functions for the Bayesian network structure learning.*

1 An introductive example

Bayesian networks are a network-based framework for representing and analyzing models involving uncertainty. They are used for *intelligent decision aids, data fusion, 3-E feature recognition, intelligent diagnostic aids, automated free text understanding, data mining*, etc. Bayesian networks came from the marriage of ideas between *artificial intelligence, decision analysis,* and *statistic communities*. Because the development of propagation algorithms followed by availability easy to use, so the number of creative Bayesian networks applications have been grown. Here is a simple example of Jan Nunnink [20] in Medical Diagnostics using Bayesian networks:

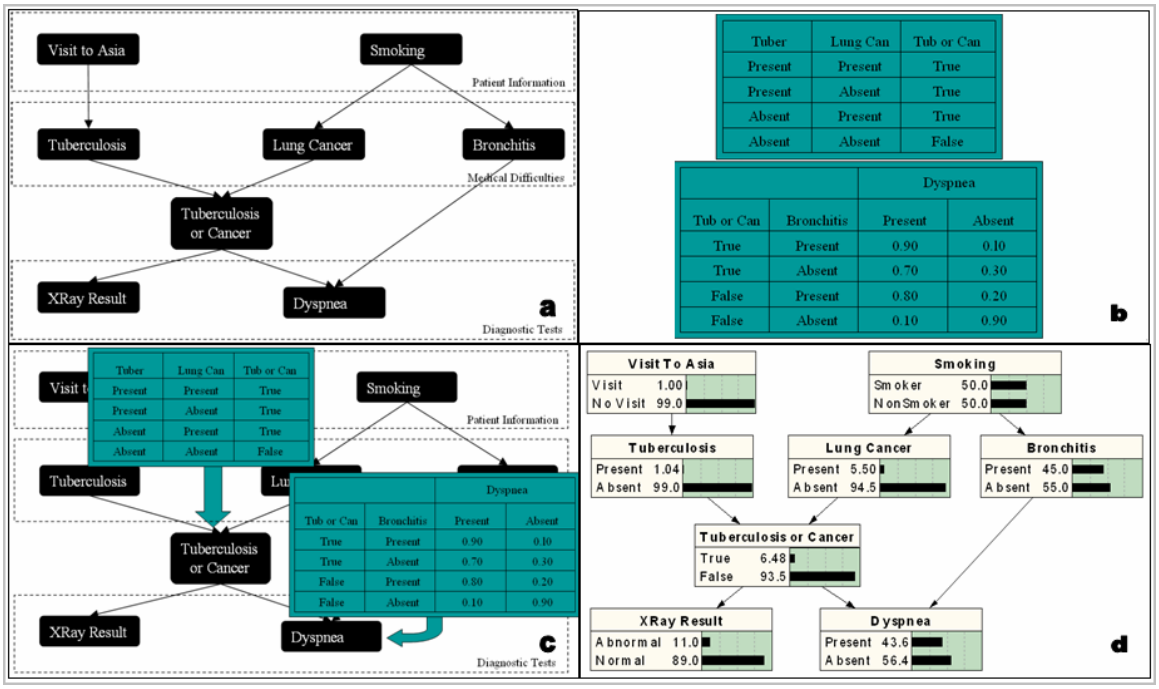

Figure 1: Example for constructing Bayesian networks from Medical Diagnostics Data

In figure 1a, network represents a knowledge structure that models the relationship between medical difficulties, their causes and effects, patient information and diagnostic tests.

In figure 1b, statistic data is collected in Medical Diagnostics.

In figure 1c, relationship knowledge is modeled by deterministic functions, logical and conditional probability distributions.

In figure 1d, propagation algorithm processes relationship information to provide an unconditional or marginal probability distribution for each node. The unconditional or marginal probability distribution is frequently called the belief function of that node.

From this network, we can use the additional information to take a prediction easily:

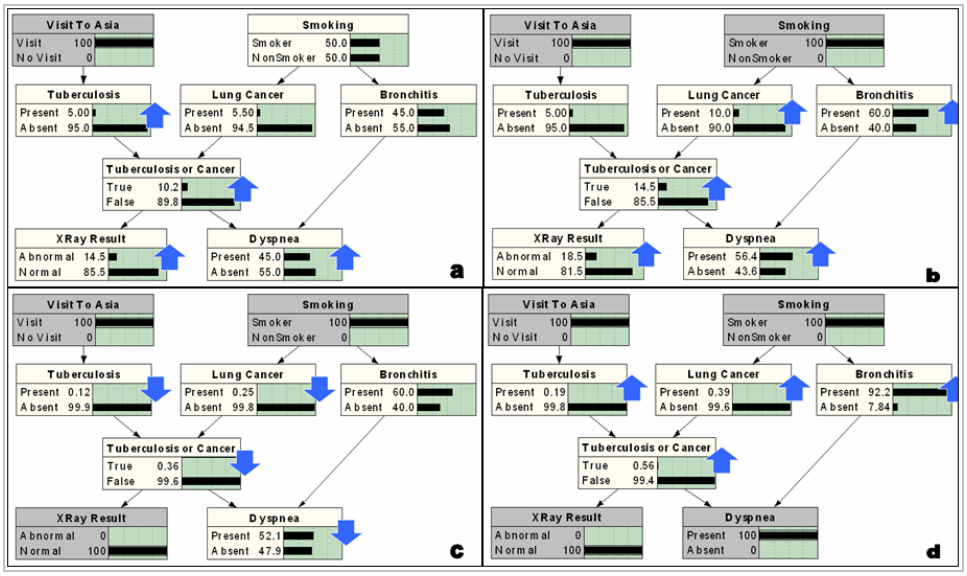

Figure 2: Example from prediction from Bayesian networks in Medical Diagnostics

In figure 2a, as a finding is entered, the propagation algorithm updates the beliefs attached to each relevant node in the network. Interviewing the patient produces the information that "Visit to Asia" is "Visit". This finding propagates through the network and the belief functions of several nodes are updated.

In figure 2b, further interviewing of the patient produces the finding "Smoking" is "Smoker". This information propagates through the network.

In figure 2c, finished with interviewing the patient, the physician begins the examination. The physician now moves to specific diagnostic tests such as an X-Ray, which results in a "Normal" finding which propagates through the network. Note that the information from this finding propagates backward and forward through the arcs.

In figure 2d, the physician also determines that the patient is having difficult breathing, the finding "Present"' is entered for "Dyspnea" and is propagated through the network. The doctor might now conclude that the patient has bronchitis and does not have tuberculosis or lung cancer.

The problem of learning a BN from a collection of observed data can be stated as follows:

The first, given the graph G, infer the parameter (local probability distribution for each node) from data. This is the problem of parameter learning/estimation. There are two approaches for this problem:

- *Frequency parameter* estimation with parameter θ is unknown *constant* (for the complete data, we use *MLE* - Maximum Likelihood Estimator, *MAP* - Maximum A Posteriori estimate; for the uncomplete data, we use *EM* - Expectation Maximization algorithm, *MCMC* - Markov chain Monte Carlo methods);

- and *Bayesian parameter* estimation with parameter θ is unknown *ramdom variable*.

The second, given a data set $D = Y_1, Y_2, Y_3, ..., Y_n$ of observed instances of a set of discrete random variables, find the most probable graph G for explaining the data contained in D. This is the problem of structural learning. There are three (3) approaches for this problem:

1. *Constraint-based learning* (also called *conditional independence tests*) [6], [9] with the testing of independencies and adding the edges according to the tests;

2. *Searching and scoring* with the definition a selection criterion that measures the goodness of a model (presented more detail in the next section);

3. and *Mixture model* (recent) [1] with the testing for almost all independencies, then searching and scoring according to the possible tests.

Searching and scoring as one common approach to this problem. It introduces a scoring metric that evaluates how probable each graph G explains the data in D. In the presence of such a scoring metric, the problem of discovering a BN then reduces to the problem of searching for a graph that yields a high score, given the observed data in D. To search the highest scoring graph, a particular search method needs to be used. The most works in Bayesian networks prefer apply the *greedy search* for this task.

This paper presents a bibliography work about the scoring functions for the Bayesian network structure learning.

2 Bayesian Scoring Functions

2.1 Notion of BN

Technically, a *static* BN [11] (*dynamic* BN for the context of time series data [19]) is an acyclic directed graph that encodes a joint probability distribution over χ, where $\chi = X_1, ..., X_n$ is a set of discrete random variables X_i. The BN for is specified as a pair $< G, \Theta >$. The variable G represents a directed graph whose vertices correspond to the random variables $X_1, ..., X_n$. In this graphical representation, each variable X_i is independent of its non-descendants given its parents in G. The variable Θ represents a set of parameters that collectively quantify the probability distributions associated with the variables in the graph. Each parameter of Θ (in a model/space) is specified by $\theta_{x_i|pa(X_i)} = P(x_i|pa(X_i))$, for each possible value x_i of X_i, and each possible value $pa(X_i)$ of $Pa(X_i)$. $Pa(X_i)$ denotes the set of parents of X_i in G and $pa(X_i)$ denotes a particular instantiation of the parents. Thus, a BN specifies a unique joint probability distribution (*parameter*) over χ given by:

$$\boxed{P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))}$$

2.2 Main idea of Scoring functions

The very *simple* idea of Bayesian scoring functions is starting from a prior probability distribution on the possible networks, then we will compute the posterior probability distribution conditioned to the available data D, $p(G|D)$. The best network is the one that maximizes the posterior probability.

The scoring functions are based on different principles, such as entropy and information [7], the minimum description length [10], [3], [4], or Bayesian approaches [15] Focusing on the methods for learning Bayesian networks based on the scoring, the problem can be formally expressed as follows: given a training data set D, find a DAG G* such that [13]:

$$\boxed{G* = arg\ \max\ Score(G:D) = logP(G|D) = logP(D|G) + logP(G) - logP(D)}$$

where $Score(G : D)$ is the scoring function measuring the degree of fitness of any candidate DAG G to the data set. According to the notion of *Likelihood*, we also have:

$$\boxed{P(D|G) = L(G, \theta_G, D) = P(D|G^h, \theta_G)}$$

where G^h: the hypothesis that the data D was generated by a distribution that can be factored according to G and θ_G is the parameter of network. However, likelihood is not useful for model selection, because of over-fitting problem. So, we have to need a posterior probability distribution conditioned to the available data D for the error estimation of model. This is the very important content of scoring function.

2.3 The state-of-the-art of Scoring functions

There are different ways to measure the degree of fitness of a DAG with respect to a data set. Most can be grouped into two categories: Bayesian and information measures. We present in the next some of typical proposed methods in the literature of model selection. For more complete information about these methods, a useful document of Martin Sewell in April 2007 [24] is recommended to "visit".

One of the first Bayesian scoring functions, called K2, was proposed by Herskovits and Cooper [14].

$$\boxed{K2(G,D) = log(p(G)) + \sum_{i=1}^{n}\left[\sum_{j=1}^{q_i}\left[log(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}) + \sum_{k=1}^{r_i} log(N_{ijk}!)\right]\right]}$$

r_i: the number of states of the variable X_i
q_i: the number of possible configurations of the parent set $Pa_G(X_i)$ of X_i
$w_{ij}, j = 1, ...q_i$, represents a configuration of $Pa_G(X_i)$;
N_{ijk} is the number of instances in the data set D where the variable X_i takes the value x_{ik} and the set of variables $Pa_G(X_i)$ take the value w_{ij};
N_{ij} is the number of instances in the data set where the variables in $Pa_G(X_i)$ take their $j-th$ configuration w_{ij}: $N_{i,j} = \sum_{k=1}^{r_i} N_{ijk}$

N_{ik} is the number of instances in D where the variable X_i takes its $k-th$ value x_{ik}: $N_{i,k} = \sum_{j=1}^{q_i} N_{ijk}$

Afterwards, the so-called BD (Bayesian Dirichlet) score was proposed by Heckerman et al. [8] as a generalization of K2:

$$\boxed{BD(G,D) = log(p(G)) + \sum_{i=1}^{n}\left[\sum_{j=1}^{q_i}\left[log(\frac{\Gamma(\eta_{ij})}{\Gamma(N_{ij}+\eta_{ij})}) + \sum_{k=1}^{r_i} log(\frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})})\right]\right]}$$

where the values η_{ijk} are the parameters for the Dirichlet prior distributions of the parameters given the network structure, and $\eta_{ij} = \sum_{k=1}^{r_i} \eta_{ijk}$. $\Gamma(.)$ is the function *Gamma*: $\Gamma(c) = \int_0^\infty e^{-u}u^{c-1}du$

It should be noted that if c is an integer, then $\Gamma(c) = (c-1)!$. If the values of all the parameters are $\eta_{ijk} = 1$, we obtain the K2 score as a particular case of BD.

In practical terms, the specification of the parameters η_{ijk} is quite difficult (except if we use non-informative assignments, as the ones employed by K2). However, by considering the additional assumption of likelihood equivalence [8], it is possible to specify the parameters relatively easily. While the result is a scoring function called BDe (and its expression is identical to the BD one in Equation of BD), the parameters can now be computed in the following way:

$$\boxed{\eta_{ijk} = \eta \times p(x_{ik}, w_{ij}|G_0),}$$

where $p(.|G_0)$ represents a probability distribution associated with a prior Bayesian network G_0 and η is a parameter representing the equivalent sample size.

A particular case of BDe which is especially interesting appears when:

$$\boxed{p(x_{ik}, w_{ij}|G_0) = \frac{1}{r_i q_i}}$$

that is the prior network assigns a uniform probability to each configuration of $X_i \bigcup Pa_G(X_i)$. The resulting score is called BDeu, which was originally proposed by Buntine [5]. This score only depends on one parameter, the equivalent sample size η, and is expressed as follows:

$$\boxed{BDe(G,D) = log(p(G)) + \sum_{i=1}^{n}\left[\sum_{j=1}^{q_i}\left[log(\frac{\Gamma(\frac{\eta}{q_i})}{\Gamma(N_{ij}+\frac{\eta}{q_i})}) + \sum_{k=1}^{r_i} log(\frac{\Gamma(N_{ijk}+\frac{\eta}{r_i q_i})}{\Gamma(\frac{\eta}{r_i q_i})})\right]\right]}$$

2.3.1 Information Criterion

Akaike [2] proposed the score AIC (the Akaike Information Criterion):

$$\boxed{AIC(G,D) = M_{ML}(G,D) - Dim(G)}$$

where $M_{ML}(G,D) = Max_{\theta_G} L(G, \theta_G, D)$

The Hannan-Quinn information criterion (HQC) [12] is an alternative to Akaike Information Criterion (AIC). It is given as:

$$\boxed{HQC(G,D) = M_{ML}(G,D) - Dim(G).log(logN)}$$

Afterward, Schwarz [23] proposed BIC (Bayesian Information Criterion). This criterion is also known as the SIC (Schwarz information criterion):

$$\boxed{BIC(G,D) = M_{ML}(G,D) - \frac{Dim(G)}{2}.logN}$$

To compare the BDe and BIC scoring metrics, by choosing to use a 3-category hard discretization of the data and a greedy search method with 1000 random restarts, [25] concluded that the BDe score works better than the BIC score in recovering the underlying simulated genetic regulatory pathway given this quantity of sampled data . The missing edges under the BIC score are consistent with its known over-penalization of model complexity. With large amounts of data, the BIC is a good approximation to the full posterior BDe score and is faster to compute; however, it is known to over-penalize with small amounts of data.

Carlos C. R. [22] and Lauría E. ([17],[18]) show that CIC (as the sum of two terms AIC, BIC) is better than its two competitors AIC and BIC:

$$\boxed{CIC(G,D) = M_{ML}(G,D) - \frac{Dim(G)}{2} - logV - \frac{\pi R}{Nlog(d+1)}}$$

where,

N: the number of observations;

$Dim(G) = 2^n - 1$;

$V = \frac{\pi^k}{(k-1)!}$, where $k = 2^{n-1}$;

$R = \frac{1}{4}d(d-1)$;

Lam and Bacchus [16] suggest as an alternative using the minimum description length (MDL) principle to bias the choice of model toward simpler ones. The minimum description length principle counsels that the best model of a collection of data items is the model that minimizes the sum of the length of the encoding of the model, and the length of the encoding of the data given the model, both of which can measured in bits.

MDL (Minimum Description Length) 1:

$$\boxed{MDL_1 = logP(G) + M_{BIC}(G,C)}$$

MDL (Minimum Description Length) 2:

$$\boxed{MDL_2 = logMDL_1 - |E_G|logN - c.Dim(G)}$$

A comparison of Ana O.B. and Martins F.V updated in 2006 [21] is recommended for a global view about scoring function. This study shows that, in the context of mixture models, AIC with penalty factors of 3 and 4, rather than the traditional value of 2, are the best segment retention criteria to use in recovering small niche segments.

Here is their 26 compared functions:

Table 1: Information Criteria and Classification Criteria

	Criteria	Description	Reference
Information Criteria		*Estimadores da distância de Kullback-Leibler*	
	Akaike Information Criteria	$AIC = -2\ln L + 2k$	Akaike (1973)
	Modified AIC 3	$AIC_3 = -2\ln L + 3k$	Bozdogan (1994)
	Modified 4	$AIC_4 = -2\ln L + 4k$	Bozdogan (1994)
	Takeuchi's Information Criterion	$TIC = -2\ln L + 2.tr\left[\mathbf{IF}^{-1}\right]$	Takeuchi (1976)
	Small sample AIC	$AIC_c = AIC + \left[2k(k+1)\right]/(N-k-1)$	Hurvich & Tsai (1989, 1995)
		Bayesian Criteria	
	Bayesian Information Criteria	$BIC = -2\ln L + k\ln N$	Schwartz (1978)
		Consistent Criteria	
	Consistent AIC	$CAIC = -2\ln L + k\left[(\ln N)+1\right]$	Bozdogan (1987)
	Information Complexity Criterion	$ICOMP = -2\ln L + k\ln\left[\frac{tr(\mathbf{F}^{-1})}{k}\right] - \ln\left\|\mathbf{F}^{-1}\right\|$	Bozdogan (1994)
	Hannan –Quinn	$HQ = -2\ln L + 2k\ln(\ln N)$	Hannan & Quinn (1979)
	Minimum Description Length 2	$MDL_2 = -2\ln L + 2k\ln N$	Liang *et al.* (1992)
	Minimum Description Length 5	$MDL_5 = -2\ln L + 5k\ln N$	Liang *et al.* (1992)
Classification Criteria		Fuzzy Indices	
	Partition Coeficient	$PC = \sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns}^2 / N$	Bezdek (1981)
	Partition Entropy	$PE = \left[\sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns}\ln p_{ns}\right]/N$	Bezdek (1981)
	Normalized Partition Entropy	$NPE = PE/\left[1 - S/N\right]$	Bezdek (1981)
	Nonfuzzy Index	$NFI = \left[S\left(\sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns}^2\right) - N\right]/\left[N(S-1)\right]$	Roubens (1978)
	Minimum Hard Tendency	$Min_{ht} = \max_{1\le s\le S}\left\{-\log_{10}(T_s)\right\}$	Rivera, *et al.* (1990)
	Mean Hard Tendency	$Mean_{ht} = \sum_{s=1}^{S} -\log_{10}(T_s)/S$	Rivera, *et al.* (1990)
		Probabilistic Indices	
	Entropy Measure	$Es = 1 - \left[\sum_{n=1}^{N}\sum_{s=1}^{S} -p_{ns}\ln p_{ns}\right]/N\ln S$	DeSarbo *et al.* (1992)
	Logarithm of the partition Probability	$LP = -\sum_{n=1}^{N}\sum_{s=1}^{S} z_{ns}\ln p_{ns}$	Biernacki (1997)
	Entropy	$E = -\sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns}\ln p_{ns}$	Biermacki (1997)
	Normalized Entropy Criterion	$NEC(s) = E(s)/\ln L(s) - \ln L(1)$	Celeux & Soromenho (1996)
	Classification Criterion	$C = -2\ln L + 2E$	Biernacki & Govaert (1997)
	Classification Likelihood Criterion	$CLC = -2\ln L + 2LP$	Biernacki & Govaert (1997)
	Approximate Weight of Evidence	$AWE = -2\ln L_c + 2k(3/2 + \ln N)$	Banfied & Raftery (1993)
	Integrated Completed Likelihood – BIC	$ICL\text{-}BIC = -2\ln L + 2LP + k\ln N$	Biernacki & Celeux (1998)
	ICL with BIC approximation	$ICOMPLBIC = -2\ln L + 2E + k\ln N$	Dias (2004)

3 Conclusion

Information theory do not enough believe for the notion of "true model". The models, by definition, are only *approximations* to unknown reality or truth; *there are no "true models" that perfectly reflect full reality.* George Box made the famous statement *"All models are wrong but some are useful"*. So, we have to need a prior distribution from the experts knowledge in a specific domain. Moreover, a "best model", for analysis of data, depends on sample size; smaller effects can often only be revealed as sample size increases.

References

[1] S. Acid and L. M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 2003. 4

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pages 716–723, 1974. 6

[3] R. R. Bouckaert. Belief networks construction using the minimum description length principle. *Lecture Notes in Computer Science*, 1993. 5

[4] R. R. Bouckaert. Bayesian belief networks: from construction to inference. *PhD thesis, University of Utrecht*, 1995. 5

[5] W. Buntine. Theory refinement of bayesian networks. *In Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, page 52, 1991. 6

[6] J. Cheng, R. Greiner, J. Kelly, D. A. Bell, and W. Liu. Learning bayesian networks from data: an information-theory based approach. artificial intelligence. *Artificial Intelligence*, 2002. 4

[7] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 1992. 5

[8] D. Geiger D. Heckerman and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 1995. 6

[9] L. M. de Campos and J. F. Huete. A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning*, 2000. 4

[10] N. Friedman and M. Goldszmidt. Learning bayesian networks with local structure. *In Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, 1996. 5

[11] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *RECOMB*, pages 127–135, 2000. 4

[12] E. J. HANNAN and B. G. QUINN. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological*, 1979. 7

[13] D. Heckerman. A tutorial on learning with bayesian networks. *Technical Report MSR-TR-95-06, Microsoft Research*, 1996. 5

[14] E. Herskovits and G. F. Cooper. Kutató: An entropy-driven system for the construction of probabilistic expert systems from databases. *In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, 1990. 5

[15] M. Kayaalp and G. F. Cooper. A bayesian network scoring metric that is based on globally uniform parameter priors. *In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 2002. 5

[16] Wai Lam and Fahiem Bacchus. Learning bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, 10(4), 1994. 7

[17] E. Lauría. Learning structure and parameters of bayesian belief networks. *The University at Albany, SUNY. School of Information Science*, 2003. 7

[18] E. Lauría. Learning the structure of a bayesian network, in maximum entropy and bayesian methods. *AIP Conf. Proc.*, 2005. 7

[19] K. Murphy and S. Mian. Modeling gene expression data using dynamic bayesian networks. *Technical Report, University of California, Berkeley*, 1999. 4

[20] Jan Nunnink. Introduction to bayesian networks. *A Tutorial for the 66th MORS Symposium*, 1998. 2

[21] Ana Oliveira-Brochado and F. Vitorino Martins. Examining the segment retention problem for the "group satellit" case. (220), July 2006. 7

[22] Carlos C. Rodríguez. The abc of model selection: Aic, bic and the new cic. *Department of Mathematics and Statistics The University at Albany, SUNY Albany, NY*, 2005. 7

[23] G. Schwarz. Estimating the dimension of a model. *Annals of Statistcs*, pages 461–464, 1978. 7

[24] Martin Sewell. Model selection. *http://www.modelselection.org/modelselection.pdf*, 2007. 5

[25] Jing Y. and V. Anne Smith. Using bayesian network inference algorithms to recover molecular genetic regulatory networks. *Duke University, Department of Electrical Engineering, Box 90291,Durham, NC 27708*, 2002. 7