Analysis of Water Chemical Contaminants: A Comparative Study

Khalil Shihab
Department of Computer Science, SQU, Box 36, Al-Khod, Oman

Abstract: Increasing the degradation of groundwater quality in Oman by salinization and chemical contaminants threaten primary sources of drinking water, especially in the coastal agricultural areas. This work elaborates the quality deterioration of groundwater due to chemical contaminants. First, we describe the development and application of Dynamic Bayesian Networks (DBNs) to determine the impact of these contaminants on groundwater quality. Second, we discuss and compare the results produced by these methods with that produced by the applications of neural networks and by the applications of classical time series models.

Keywords: Bayesian Reasoning, SOM, Groundwater Quality Assessment, Classical Time Series.

1 Introduction

Water is an essential requirement for irrigated agriculture, domestic uses, including drinking, cooking and sanitation, as critical input in industry. Declining surface and groundwater quality is regarded as the most serious and persistent issue affecting Oman in particular. The Sultanate faces severe challenges as it confronts the extremely growing and complicated issues of contamination of the groundwater supply in and around hazardous waste disposal sites across the nation. In Salalah area of Oman, groundwater has been an important natural resource and the only available water source other than the seasonal rainfall.

Groundwater quality and pollution are determined and measured by comparing physical, chemical, biological, microbiological, and radiological quantities and parameters to a set of standards and criteria. A criterion is basically a scientific quantity upon which a judgment can be based. In this work, however, we considered only the chemical parameters, total dissolved solids (TDS), electrical conductivity (EC) and water pH, section 4 presents more details.

Various countries have attempted to develop satisfactory procedures for assessing, monitoring and controlling contamination of the groundwater supply in and around hazardous waste disposal sites [1]. These attempts resulted in various environmental regulations that focus attention on the maximum allowable limits of hazardous pollutants in the groundwater supply. However, they pay scant attention to the nature of groundwater data and the development of valid statistical procedures for detecting and monitoring groundwater contamination.

Recent attempts based on Artificial Intelligence (AI) were first applied to the interpretation of biomonitoring data [8]. Other works were based on pattern recognition using artificial neural networks (NNs). A more recent study described a prototype Bayesian belief network for the diagnosis of acidification in Welsh rivers. Hobbs [5] uses Bayesian probabilities to examine the risk of climate change on water resources.

2 Problem Description

Oman, has very substantial groundwater resources on which the country's agriculture depends. The oil boom, the resultant population boom (possibly fivefold since the 1960's) and the new investment have led to a large expansion in irrigated areas.

The Salalah plain extends over a 253 km2 area to the north of the Omani coastline of the Arabian Sea to the Mountains of Dhofar.

In this work, we compare the applications of Bayesian techniques, classical statistical analysis, and SOM to forecast groundwater pollution levels in the Salalah plain, see Figures 1.

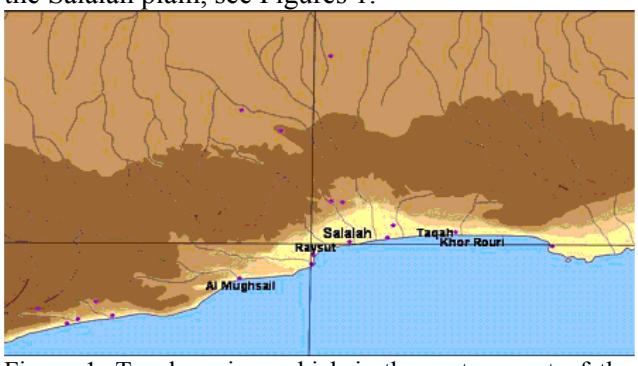

Figure 1. Taqah region, which is the eastern part of the Salalah plain in Oman.

3 Data Collection

The Ministry of Water Resources (MWR) maintains data on the concentration of the harmful substances in the groundwater at Taqah monitoring sites, which are located to the south of the Sultanate of Oman, in the Salalah plain [9]. The MWR identified that the datasets collected from these monitoring wells in the Sultanate are important in assessing the groundwater quality and in the prediction of the effect of certain pollutants on drinking water. The period covered in these locations is from 1984 to 2004 [3, 4]. Each site has several monitoring wells and water samples were collected periodically from these wells and the concentration of the pollutants in these water samples was recorded. We also collected data for the period 1984-1994 from Oman Mining Company (OMCO) and Ministry of Environmental and Regional Municipalities (MRME). However, the datasets are not complete. We, therefore, filled the gaps with data collected by some researchers at the Sultan Qaboos University.

4 Dynamic Bayesian Networks (DBNs)

The problem of assessing and forecasting water quality requires not only modelling the static probabilistic dependencies between its constituents but also the dynamic behaviour of these constituents. Dynamic Bayesian Networks (DBNs) can easily capture these static and dynamic behaviours [4]. They extend Bayesian Networks from static domains to dynamic domains [1]. A static Bayesian Network can be extended to a Dynamic Belief Network by introducing relevant temporal dependencies between the representations of the static network at different times. In contrast to the time series models that use regression to represent correlations, DBNs represent the temporal causal relationships between variables. Therefore, DBNs can introduce more general dependency models [5].

DBNs can be effectively and cheaply used for monitoring and predicting complex situations. For example, they have been used for monitoring and controlling highway traffic for identifying gene regularity from microarray data [9], and for prediction of river and lake water pollution.

The temporal repetition of identical model structures encourages the integration of object oriented techniques with Bayesian networks. It started with methods for reusing elements of network specifications and division of large networks into smaller pieces. These and other successful object-oriented Bayesian networks (OOBNs) models and their applications to real-world problems have greatly encouraged us to develop a model and a computer system based on the OOBN representation to assess and predict the water quality. Therefore, we used the Hugin and dHugin tools for implementing our Bayesian networks [6]. The Hugin system allows the implementation of an OOBN. The system considers a Bayesian Network (BN) as a special case, initial building network, of an OOBN. Other networks in the OOBN are nodes that represent instances of the base network.

4.1 Bayesian Networks Development

Identifying the domain variables (pollution constituents) and the causal relationships between these variables constitute the main part of development process. In our study, we only considered the dependencies between total dissolved solids (TDS), electrical conductivity (EC) and water pH. In the Sultanate of Oman, these are the main factors that industry experts were dealing with and, therefore, maintaining good data about them. In fact, we used our literature-based network structure as a starting point for discussion with the experts to explain the Bayesian network approach and to get their input [1]. In addition, we analyzed the data collected from many wells and the results revealed that these chemical parameters are useful indicators of groundwater quality because they form the majority of the variance in the data scatter.

The electrical conductivity (EC) of the water has been used as a measure for the salinity hazard of the groundwater used for irrigation in the Salalah plain. According to international water-quality standards, irrigation water with EC values up to 1 mS/cm (where mS/cm = milli-Siemens per centimeter) is safe for all crops and between 1 and 3 mS/cm is acceptable, but values higher than 3 mS/cm restrict the use of water for many irrigated crops. Changes in conductivity can be caused by changes in water content of the soil and by soil or groundwater contamination.

The total dissolved solid (TDS) limit is 600 mg/L, which is the objective of the current plan of the MWR. TDS contains several dissolved solids but 90% of its concentration is made up of six constituents. These are: sodium Na, magnesium Mg, calcium Ca, chloride Cl, bicarbonate HCO_3 and sulfate SO_4. We, therefore, considered only these elements in the calculation of TDS. Other factors that are considered less significant to groundwater quality in Oman were not recoded and therefore neglected in this study.

Both TDS and EC can affect water acidity or water pH. Solute chemical constituents are variable

in high concentration at lower pH (higher acidity). On the other hand, acidity allows migration of hydrogen ions (H+), which is an indication of conductivity. Therefore, our work concentrated on the following relations.

TDS → EC, EC → pH, TDS → pH

Reaching to these relations we used two learning approaches to construct and parameterize a simple static BN that have three nodes, each node represents a groundwater quality constituent (TDS, EC or pH). Learning basically consists of two different components: 1) learning the network structure , 2) learning the conditional probability distributions.

For the first approach, we used the Hugin system that supports structure and parameter learning in Bayesian networks. Regarding the former, the PC learning algorithm was applied, while for the latter, we used the EM algorithm.

In the second approach, we used a program written in C++ to generate the conditional probabilities. The program requires, as inputs, a dataset and thresholds or discretization intervals for each variable in the dataset. It calculates the frequencies of various values of the variables given various values of their parents in the network. These frequencies estimate the conditional probability tables associated with each node. Knowing that the maximum allowable TDS, EC and pH in the drinking water are 550 mg/l, 670 mg/l and 7.5 respectively, we pass these values as thresholds along with the dataset to our C++ program. For TDS, the program divides the relevant dataset into two categories, considering TDS=550 as a threshold. Thus, the first category has TDS < 550 and the second category has TDS >= 550. For EC, the program also divides the data sample into two categories: data with EC < 670 and data with EC >= 670. Finally, for pH, the program also divides the relevant dataset into two categories, data with pH < 7.5 and data with pH >= 7.5.

After the categorization of the dataset, the program uses the following algorithm to produce the conditional probabilities.

Let A and B be two events (for example A=EC<670) and B=TDS<550, and let m(A) and m(B) be the frequencies of A and B. The program calculates P(A/B) as follows:

```
if A includes B then
    P(A/B)=1;
else
    X=A∩B;
 if X=Φ then
    P(A/B)=0;
 else
```

$$P(A/B) = \frac{\sum_{x \in X} m(X)}{m(B)};$$

According to international water-quality standards, irrigation water with EC values up to 1 mS/cm is safe for all crops and between 1 and 3 mS/cm is acceptable, but values higher than 3 mS/cm restrict the use of water for many irrigated crops. Changes in conductivity can be caused by changes in water content of the soil and by soil or groundwater contamination.

The total dissolve solid (TDS) limit is 600 mg/L, which is the objective of the current Plan of the MWR. TDS contains several dissolved solids but 90% of its concentration is made up of six constituents. These are: sodium Na, magnesium Mg, calcium Ca, chloride Cl, bicarbonate HCO3 and sulfate SO4. We, therefore, considered only these elements in the calculation of TDS, which is represented as a node without parents in the network structure. This simplification is necessary to make the problem tractable and to keep it consistent with available data without losing information.

Both TDS and EC can affect water acidity or water pH. Solute chemical constituents are variable in high concentration at lower pH (higher acidity). On the other hand, acidity allows migration of hydrogen ions (H+), which is an indication of conductivity. Therefore, our work concentrated on the following relations:

TDS → EC,
EC → pH,
TDS → pH.

Table 1. TDS data for the well Well 001/577.

Yr	Mg	SO_4	Na	Ca	K	Cl	HCO_3
84	12	11	21	91	11	172	224.7
85	10	12	20	88	13	148	234.5
86	9	14	18	92	17	140	275.4
87	14	12	43	86	14	148	287.2
88	12	12	20	90	20	132	255.8
89	10	12	17	86	16	148	276.9
90	32	13	15	92	18	164	224.6
91	19	11	45	89	21	168	287.4
92	12	14	152	92	17	176	291.5
93	21	12	165	93	19	192	296.7
94	27	14	88	96	23	204	294.4
95	7	11	65	60	22	140	310.8
96	16	25	64	52	15	244	321.5

97	13	19	83	102	18	204	314.6
98	19	26	97	107	26	248	412.6
99	56	38	217	98	57	220	487.7
00	41	20	201	104	31	236	388.4
01	43	23	204	135	30	244	387.6
02	55	32	210	138	41	308	438.6
03	52	20	147	121	33	272	410
04	48	21	152	130	34	284	405.4

Table 2. TDS, EC, and pH data for the well Well 001/577.

Yr	TDS	EC	pH
84	543	548	7.7
85	526	548	7.8
86	565	579	7.75
87	604	588	7.57
88	542	601	7.43
89	566	625	7.34
90	559	638	7.32
91	640	798	7.27
92	755	739	7.24
93	799	758	7.28
94	746	799	7.29
95	616	514	7.3
96	738	619	7.28
97	754	869	7.19
98	936	558	7.15
99	1174	855	7.15
0	1021	796	7.06
1	1067	855	6.98
2	1223	844	6.94
3	1055	881	6.9

Table 1 shows the monitoring measurements of the main components of TDS for the well Well 001/577. Data for the constituents Mg, SO4, Na, Ca, K, and CL of TDS were only reported. Differences for other parameters were not significant in the Salalah area. Table 2 shows the measurements of EC and pH for the same well along with the measurements of TDS copied from Table 1.

After providing the prior probabilities and the conditional probability tables, the results of the run session for new-presented data for any selected node of HUGIN are also shown in Figures 2 and 3.

Once the static BN model (static model) for each monitoring well was built, parameterized and tested, we used these models as initial building networks in the construction of OOBN for groundwater quality prediction. The OOBN, as shown in Figure 2, models the time slices for each well characterizing the temporal nature of identical model structures, where the initial building network, see Figure 3, describes a generic time-sliced network.

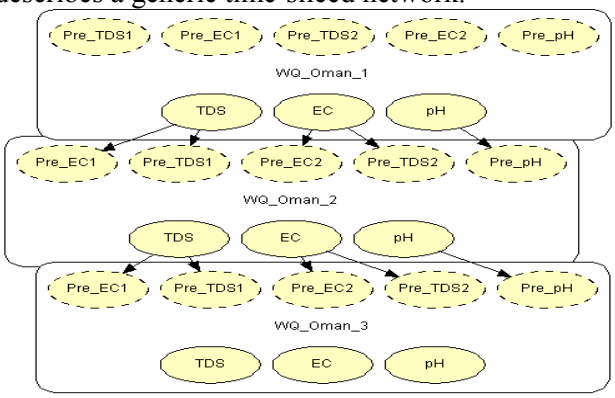

Figure 2. The OOBN representing three time-sliced network.

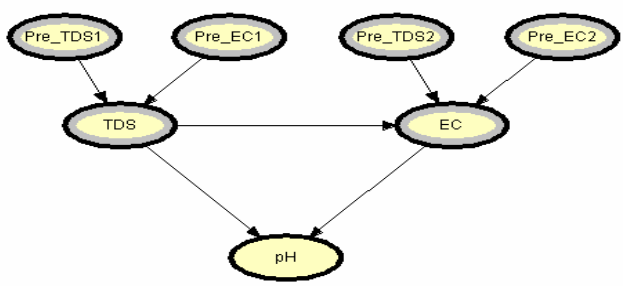

Figure 3. The initial building block representing one time-sliced network.

5 Application Results

The application was carried out with special emphasis on the advantages of Bayesian against traditional techniques. It involved monitoring the groundwater quality parameters and validating crucial assumptions.

Figures 4 show the KL-divergence between the true and the approximate distribution. Since the KL-distance converges to zero, this is an indication of the accuracy and reliability of OOBN.

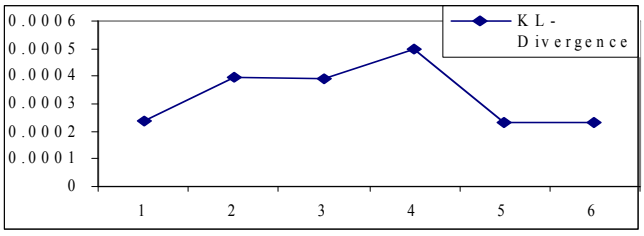

Figure 4. The KL-divergence between the true distribution and the estimate distribution over all variables.

6 Using Classical Time Series

We applied the classical time series analysis to groundwater quality data and to compare the results

with that obtained by the application of Dynamic Bayesian Networks (DBN).

Time series analyses of water supply wells with respect to the concentration of chemical constituents are presented in Figures 5-6.

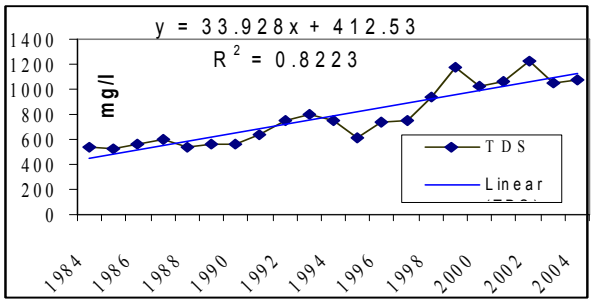

Figure 5. Fluctuation of TDS concentration for the well Well001/577.

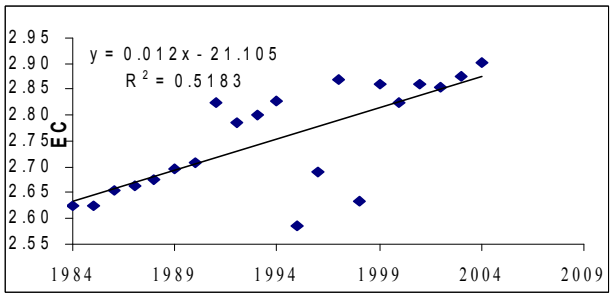

Figure 6. EC concentration is poorly represented for the well Well001/577.

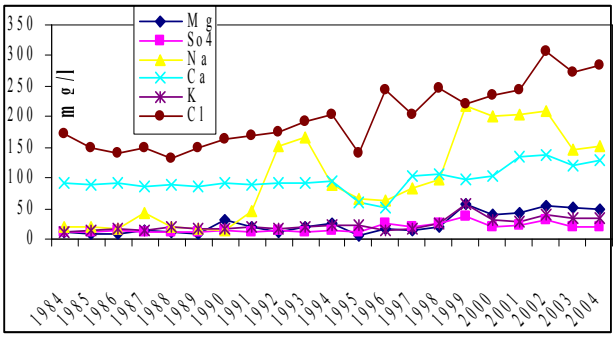

Figure 7. Fluctuation of the concentration of the major chemical constituents for Well001/577.

The fluctuation of the concentration of the chloride (Cl), sodium (Na), and calcium (Ca) with respect to time is shown in Figure 7. The values were averaged during the initial analysis as there were no significant differences among the monthly data. Chloride values above 250 mg/l give a slight salty taste to water which is objectionable by many people. Multiple regression analysis is used to explain as much variation observed in the response variable as possible, while minimizing unexplained variation from "noise". The results of this analysis is used to produce the moving average chart, Figure 9.

We used Excel Business Tools, Microsoft Excel, and Matlab for producing these and other charts.

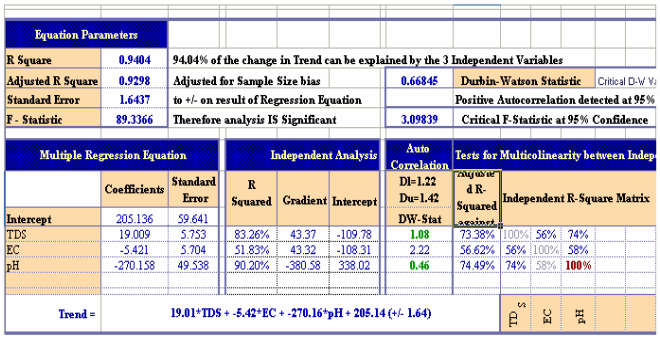

Figure 8. Excel templates for financial analysis and business productivity from Excel Business Tools .

As it is shown in Figure 9 that the trend is as follows:

TrendWQ=19.01*TDS - 5.42*EC -270.16*pH + 205.14

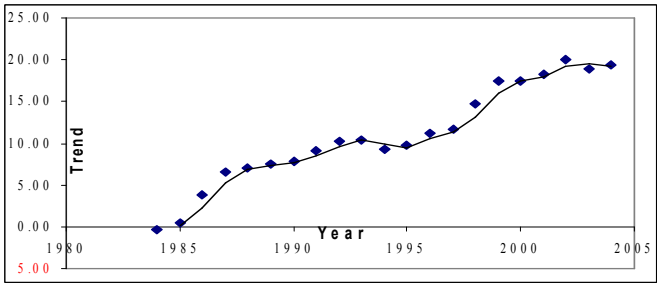

Figure 9. Moving average chart of 2-year period for groundwater quality trend.

The classical time series models are based on the assumption of stationary. Therefore, these models do not readily adapt to domains with dynamically changing model characteristics, as is the case with groundwater quality assessment. In addition, the classical models are restricted in their ability to represent the general probabilistic dependency among the domain variables and they fail to incorporate prior knowledge.

The observed groundwater quality data are irregularly spaced and not predetermined as in the case with ordinary time series. This may cause the traditional time series techniques to be ineffective (Prediction: what is the predicted value for one period ahead). It is evident that the time series casts doubts on the positive or negative effects of any chemical constituent on the groundwater quality for the long run, and is thus not as clear and reliable as in the case of using Dynamic Bayesian Techniques. While some groundwater quality constituents, such as chloride and TDS, show an increasing trend, the other constituents, such as pH, Mg, and SO4 do not demonstrate obvious trends. Therefore, we cannot draw a reliable conclusion on the cause of the

increasing trend of the groundwater quality. In addition to the ignorance of the cause-effect relationships, classical time series models assume the linearity in the relationships among variables and normality of their probability distributions.

7 Using SOM

In this section, we explain how neural methods might be applied to groundwater quality assessment.

SOM uses an unsupervised approach to learning. The map resembles a landscape in which it is possible to identify borders that define different clusters. With this map we wanted to see how different chemical values are situated in comparison to each other and to the previous years' values. It is visualized in order to discover how the groundwater quality has been changed according to different years. On the map, we define the clusters by looking at the color shades of the borders between the neurons (nodes). The dark colors in the walls represent great distances while brighter colors indicate similarities amongst the neurons. The colored borders between the nodes are of great value when trying to determine and interpret clusters.

A number of clusters and the characteristics of these clusters were identified, see Figure 10. There is a tendency starting from the bottom right corner, where the early eighties data are allocated, up towards the top of the map and then to the bottom left corner. Therefore, the map shows that the degradation of groundwater quality is in the bottom left corner where the data for the years 2000 to 2004 are located. The first seven years data from 1984 to 90 inclusive are in the bottom right corner of the map. The 1998 data are also at the ultimate bottom right of the map that shows a significant improvement in the groundwater quality for this year.

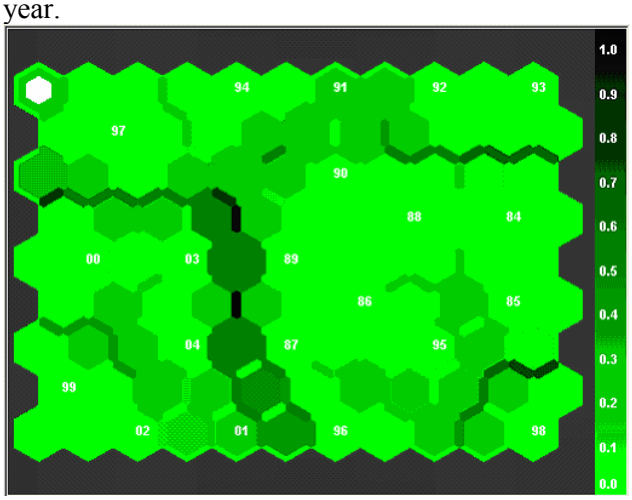

Figure 10: The u-matrix (unified distance matrix) visualization of the SOM for the chemical constituents dataset in Table 1. The map is 11 × 9 neurons. The map is also labeled.

8 Conclusion and Further Work

The simple Bayesian network presented here is the first step towards having a comprehensive network that contains the other variables that are considered by the researchers significant for the assessment of groundwater quality in the Salalah plain in particular. These variables include:

- NO3: Nitrate is an increasingly important indicator of water pollution.
- COD: chemical oxygen demand, it reflects the organic and inorganic content of the water.

References:

[1] Borsuk, M et al. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis, *Ecological Modeling*,,173, 2004, pp. 219–239.

[2] Brandherm, B. and Jameson, A. An extension of the differential approach for Bayesian network inference to dynamic Bayesian networks, *International Journal of Intelligent Systems*, 19(8), 2004, pp. 727–748.

[3] Dames and Moore. Investigation of The Quality of Groundwater Abstracted from the Salalah Plain: Dhofar Municipality, Final Report, 1992.

[4] Entec Europe Limited. Consultancy Services for The Study of Development Activities on Groundwater Quality of Salalah Well field, 1998.

[5] Hobbs B. F. Bayesian methods for analyzing climate change and water resource uncertainties, *Journal of Environmental Management*, 49, 1997, pp. 53-72.

[6] HUGIN Expert Brochure. HUGIN Expert Denmark, (http://www.hugin.com), 2007.

[7] Kevin, B. and Nicholson, *A. Bayesian Artificial Intelligence*, Chapman & Hall/CRC, 2004.

[8] Kim, S. et al. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*, 75, 2004, pp. 57–65.

[9] Ministry of Water Resources (MWR), Sultanate of Oman, 2004.

[10] Zou, M. and Conzen, S. D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data, *Bioinformatics*, 21(1), 2005, pp. 71–79.