Application of Bayesian networks and data mining to biomedical problems

Alla R. Kammerdiner*, Anatoliy M. Gupal† and Panos M. Pardalos*

*Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611, USA.
†Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine

Abstract. During the last several decades, the Bayesian networks have turned into a dynamic area of research. This great interest is owning to the advantages offered by special structure of Bayesian networks, which allows them to be very efficient in modeling domains with inherent uncertainty. Bayesian networks techniques can be successfully applied to mining various types of biomedical data.

This chapter demonstrates how various complex research problems in biology, biomedicine and other fields can be solved by means of the combination of methods from Bayesian networks and data mining.

Keywords: naïve Bayes, Bayesian networks, classification, neuroregeneration model, human genome

PACS: 01.10.Fv

1. INTRODUCTION

The foundation for Bayesian networks comes from statistics. The first formal introduction of the Bayesian networks in the field of artificial intelligence and expert systems can be traced to the works of by Pearl [1] in 1982, and Spiegelhalter and Knill-Jones [2] in 1984. The first practical applications of Bayesian networks, Munin [3] and Pathfinder [4], appeared in 1989 and 1992, respectively. Since then, the scope of research in Bayesian networks has increased dramatically. In particular, various modifications of Bayesian networks has been used for solving many challenging problems in data mining, pattern recognition, image processing and data fusion, engineering, etc.

Mathematically, a Bayesian networks on a finite collection of random variables \( X = \{X_1, X_2, \ldots, X_n\} \), can be defined as an ordered pair \((G,D)\), where

- \( G \) is a graphical component represented by a directed acyclic graph with nodes corresponding to the random variables \( X_1, X_2, \ldots, X_n \), and directed arcs symbolizing conditional dependencies between the variables. The set of all the arcs of \( G \) satisfies the following assumption: Each random variable in the graph is conditionally independent of its non-descendants in \( G \), given its parents in \( G \).
- \( D \) is a probabilistic component representing the set of parameters that, for each variable \( X_i \), \( 1 \leq i \leq n \), define the conditional distribution of \( X_i \) given its parents in the graph \( G \).
A complex concept of a Bayesian network is based on a simple statistical formula called Bayes rule:

\[ Pr(B | A) = \frac{Pr(A | B) \cdot Pr(B)}{Pr(A)} , \quad \forall \text{events } A, B \]  

(1)

which allows for probability updating, i.e. calculating the posterior probability from data given reasonable assumptions about prior probabilities.

For a comprehensive review of the relative definitions, cases and examples of Bayesian networks, as well as the methods and applications of Bayesian networks see, for example, \([5, 6]\).

Bayesian networks offer a very efficient way of modeling situations with inherent uncertainty. In many such situations, the underlying processes are mostly unclear, and the task is to try extracting the information behind such processes from the observed data. Luckily, a rapid technological progress in data collection and storage allows collecting a vast amount of data, and the data mining techniques can greatly facilitate prompt analysis and summarization of massive data. All the above is especially relevant when solving problems in biomedicine. For example, in numerous biomedical problems, one has to perform microarray DNA experiments in order to elucidate the underlying biological processes that are not yet completely understood. It is natural to apply data mining to handle the massive datasets that arise in microarray experiments. In general, it is plausible to try modeling the situation by means of Bayesian networks, when the situation exhibits causality and uncertainty, which is often the case in biological processes.

The chapter is organized as follows. Section 2 discusses the broad scope of the diverse areas of applications of Bayesian network methods, and then presents some illustrative examples of recent applications of Bayesian networks to various problems in expert systems and data mining. Section 3 continues with focuses exclusively on the application of Bayesian network to data mining, and the connection between them. Section 4 presents applications to two important biomedical problems. The chapter is concluded by the summary in Section 5.

2. ON THE BREADTH OF APPLICATIONS OF BAYESIAN NETWORKS

A concept of Bayesian network can be interpreted in different contexts. From a statistical point of view, a Bayesian network can be defined as a compact representation of the joint probability over a given set of variables. From a broader point of view, a Bayesian network is a special type of a graphical model capable of reflecting causality, as well as updating its beliefs in view of received evidence. All these features make a Bayesian network a versatile instrument that can be used for various purposes, including facilitating communication between human and computer, extracting hidden information and patterns from data, simplifying decision making, etc.

Due to their special structure, Bayesian networks have found many applications in various areas such as artificial intelligence and expert systems, machine learning and data mining. Bayesian networks are used for modeling knowledge in text analysis, image processing, speech pattern analysis, data fusion, engineering, biomedicine, gene and protein
regulatory networks, and even meteorology. Furthermore, it has been expressed that the inductive inference procedures based on Bayesian networks can be used to introduce inductive reasoning in such previously a strictly deductive science as mathematics [18]. The large scope of different application of Bayesian networks becomes especially impressive taking into account that Bayesian networks is a somewhat new theory that has been around approximately for a mere quarter of a century. Next, several examples of the recent real-life applications of Bayesian networks are considered to illustrate the above point.

Recent research in the field of automatic speech recognition [7] indicates that dynamic Bayesian networks can effectively model hidden features in speech including articulatory and other phonological features. Both hidden Markov models (HMM), which are special case of dynamic Bayesian networks (DBN), and more general dynamic Bayesian networks have been applied for modeling audio-visual speech recognition. In particular, a paper by A.V. Nefian et al. [8] describes an application of the coupled HMM and the factorial HMM as two suitable statistical models for audio-video integration. The factorial HMM is a generalization of HMM, where the hidden state is represented by a collection of variables also called factors. These factors, although independent of each other, all impact the observations, and hence become connected indirectly. The coupled HMM is a DBN represented as two regular HMM whose hidden state nodes have links to the hidden state nodes from the next time slice. The coupled HMM has been also applied to model hand gestures, the interaction between speech and hand gestures, etc. In addition, face detection and recognition problems have been studied with help of Bayesian networks.

Note that different fields of application may call for specialized employment of Bayesian network methods, and conversely, similar approaches can be successfully used in different application areas. For instance, along with the applications to speech recognition above, the coupled hidden Markov models have been employed in modeling multi-channel EGG data [19].

An interesting example of Bayesian network in application to expert systems includes developing strategies for troubleshooting complex electro-mechanical systems presented in [9]. The constructed Bayesian network has the structure of naïve Bayes model. In the decision tree for the troubleshooting model, the utility function is given by the cost of repair. Hence, the goal is to find a strategy minimizing the expected cost of repair.

An interesting recent study [10] describes some applications of Bayesian networks in meteorology from data mining point of view. A large data base of daily observations of precipitation levels and maximum wind speed is collected. The Bayesian network structure is constructed from meteorological data by using various approaches, including batch learning procedure and simulation techniques. In addition, an important data mining application of Bayesian networks is illustrated by giving an example of missing data values estimation from the evidence received.
3. APPLICATIONS OF BAYESIAN NETWORKS TO DATA MINING

Rapid progress in data collection techniques and data storage has enabled an accumulation of huge amounts of experimental, observational and operational data. As the result, massive data sets containing gigantic bulk of information can be found almost everywhere. A well-known example of a massive data set is the data set containing the observed information about human genome. The tremendous need to quickly and correctly analyze or manipulate such enormous data sets facilitated the development of data mining techniques.

Data mining is the research aimed at discovery of various types of knowledge from the large data warehouses. Data mining can also be seen as an integral part of more general process of knowledge discovery in databases. The other two parts of the knowledge discovery in databases are the preprocessing, and postprocessing. As seen above Bayesian networks can also extract knowledge from data, which is called evidence in the Bayesian framework. In fact, the Bayesian network techniques can be applied to solve data mining problems, in particular, classification.

3.1. Application of naïve Bayes to data mining

Many effective techniques in data mining utilize the methods from other multidisciplinary research areas such as database systems, pattern recognition, machine, learning, and statistics. Many of these areas have a close connection to Bayesian networks. In actuality, data mining is utilizing a special case of Bayesian networks, namely, naïve Bayes, for performing effective classification. In a data mining context, a classification is a task of assigning objects to their relevant categories. The incentive for performing classification of data is to attain a comprehensive understanding of differences and similarities between the objects in different classes.

In the Bayesian framework, the data mining classification problem translates into finding the class parameter $y$ maximizing the posterior probability of the unknown instance. This statement is also called the maximum a posteriori principle, or abbreviated, MAP. Let us show how the naïve Bayes can be represented by the following Bayesian network model:

The graph of the naïve Bayes model consists of one hypothesis variable, say $Y$, and a finite set of information variables $X_1, X_2, \ldots, X_n$ with the arcs from $Y$ to every $X_k$, $1 \leq k \leq n$. In other words, the variables form a diverging connection with the information variables $X_1, X_2, \ldots, X_n$ having a common parent $Y$. The probability distributions are given by the values $Pr(X_k \mid Y)$, for each information variable $X_k$, $1 \leq k \leq n$.

Let us apply this simple Bayesian network model to construct a so called naïve Bayes classifier, which can be used for estimating the posterior probabilities of data having discrete and continuous parameters. Let $x_1, x_2, \ldots, x_n$ denote the known parameter values, and let $y$ to be an unknown parameter. In the context of the naïve Bayes model, $x_1, x_2, \ldots, x_n$ are some realizations of information variables $X_1, X_2, \ldots, X_n$. In other words, we received evidence on $X_1, X_2, \ldots, X_n$, which results in the hypothesis vari-
able \( Y \) being in the unknown state \( y \). The posterior probability of \( y \) given \( x_1, x_2, \ldots, x_n \) can be calculated according to the Bayes rule (1) as follows:

\[
Pr(y \mid x_1, x_2, \ldots, x_n) = \frac{Pr(x_1, x_2, \ldots, x_n \mid y) \cdot Pr(y)}{Pr(x_1, x_2, \ldots, x_n)}.
\] (2)

Next, from the naïve Bayes model, we can draw additional information about the nature of the relationships among parameters. In particular, since the variables \( X_1, X_2, \ldots, X_n \) and \( Y \) form a divergent connection with \( Y \) as the parent, then from the \( d \)-separation rule, \( X_1, X_2, \ldots, X_n \) are \( d \)-separated, given \( Y \) received an evidence. That is, \( X_1, X_2, \ldots, X_n \) are conditionally independent of each other, given \( Y \). To put this mathematically,

\[
Pr(x_i, x_j \mid y) = Pr(x_i \mid y) \cdot Pr(x_j \mid y),
\] (3)

for all \( i, j \), where \( 1 \leq i, j \leq n \). Hence, the naïve Bayes model implies the following:

\[
Pr(x_1, x_2, \ldots, x_n \mid y) = \prod_{i=1}^{n} Pr(x_i \mid y).
\] (4)

Then, from (2) it follows that

\[
Pr(y \mid x_1, x_2, \ldots, x_n) = \frac{\prod_{i=1}^{n} Pr(x_i \mid y) \cdot Pr(y)}{Pr(x_1, x_2, \ldots, x_n)}.
\] (5)

Therefore, the task is to find such value of parameter \( y \) that maximizes the expression in the left-hand side of formula (5). Note that the denominator \( Pr(x_1, x_2, \ldots, x_n) \) can simply be ignored, since it does not change with the change in the value of parameter \( y \), and so, does not affect the maximum. Hence, it is sufficient to find \( y \) that maximizes the product \( \prod_{i=1}^{n} Pr(x_i \mid y) \cdot Pr(y) \).

Thus, the above classification method based on the naïve Bayes model, works in the following fashion. To classify an unknown sample \((x_1, x_2, \ldots, x_n, y)\) with unknown \( y \) and known values \( x_1, x_2, \ldots, x_n \), the naïve Bayes classifier

- computes the product \( \prod_{i=1}^{n} Pr(x_i \mid y) \cdot Pr(y) \), for each value of unknown parameter \( y \);
- selects the parameter \( y_\ast \) such that

\[
\prod_{i=1}^{n} Pr(x_i \mid y_\ast) \cdot Pr(y_\ast) = \max_y \prod_{i=1}^{n} Pr(x_i \mid y) \cdot Pr(y).
\] (6)

Now consider some common approaches to estimation of probabilities \( Pr(y) \) and \( Pr(x_i \mid y) \), where \( 1 \leq i \leq n \). First, note that as a class label parameter \( y \) must necessarily follow a discrete distribution. Hence, the probability \( P(y) \) can be estimated as the frequency of instances that belong to class \( y \) (i.e. instances with \( Y = y \)) in the training sample \( D \). On the other hand, each of parameters \( x_1, x_2, \ldots, x_n \) representing the known information, may follow either continuous or discrete distribution.

For discrete parameters \( x_i \), the conditional probabilities \( Pr(x_i \mid y) \) can be estimated similarly to \( Pr(y) \) above. In particular, \( Pr(x_i \mid y) \) is calculated approximately as the
fraction of such instances in class \( y \), where \( X_i = x_i \). Unfortunately, the described method of the probability approximation as the corresponding frequency is known to have a potential disadvantage. Indeed, if the training sample \( D \) does not contain an observation for some \( x_i \), then the above estimation approach gives \( \Pr(x_i) = 0 \). When the estimated probability equals zero for even a single parameter, the entire posterior probability involving this parameter becomes zero. One of the commonly used ways to resolve this problem is by using the \textit{Laplace correction} to obtain a non-zero estimate:

\[
\Pr(x_i \mid y) = \frac{n_i + m \cdot p}{n + m},
\]

where \( n \) represents the total number of instances of class \( y \) in the training dataset \( D \), \( n_i \) is the number of instances in \( D \) from class \( y \) for which \( X_i = x_i \), positive real \( m \) is a parameter often called the equivalent sample size, and \( 0 < p < 1 \) is some specified prior probability.

For continuous parameters, the common approach to estimation of its conditional probabilities is to make a simplifying assumption about the form of probability distribution. Usually, a continuous parameter is approximated by a normal distribution with some parameters \( \mu \) and \( \sigma \). Both the mean of the distribution, \( \mu = \mu(y) \), and its standard deviation \( \sigma = \sigma(y) \) depend on the class label \( y \), and so, we have:

\[
\Pr(X_i = x_i \mid Y = y) = \frac{1}{\sqrt{2\pi\sigma_i(y)}} \exp \left\{ -\frac{(x_i - \mu_i(y))^2}{2\sigma_i^2(y)} \right\},
\]

where \( \mu_i(y) \) and \( \sigma_i(y) \) are the mean and the standard deviation of \( x_i \), respectively.

The classification techniques based on the naïve Bayes model have the following positive characteristics:

- When the amount of missing data is insignificant, Naïve Bayes classifiers can easily handle such data by ignoring the instance during the estimations;
- If a statistically small amount of noisy data is present, naïve Bayes classifiers are robust to the noise, since the noise values is averaged out in the estimations of probability from data;
- For the same reasons, naïve Bayes classifiers are robust to irrelevant attributes.

### 3.2. Using Bayesian networks for classification

In spite of all the positive characteristics of the naïve Bayes classifier, there is a certain drawback. In general, for many practical data sets, most of the parameters are not totally independent of each other. Hence, the independence assumption, which is essential for the Naïve Bayes classifier to be exact, may not hold. However, this issue can be successfully resolved by designing a suitable Bayesian network for each individual case. Clearly, the simple Bayes classifier can be generalized into a classification method based on designing a specialized Bayesian network, whose structure appropriately models conditional independence assumptions among parameters.
Let $x$ denote the known parameters, and $y$ be unknown class label. The approach to building Bayesian networks classifiers can be summarized as follows:

- Test the parameters to find the conditional independence structure;
- Design a Bayesian network $M$ that correctly reflects the conditional independence structure;
- Estimate $Pr(x \mid y)$ from Bayesian network $M$ by using a suitable probability updating technique;
- Select parameter $y$ maximizing $Pr(y)Pr(x \mid y)$

Similarly to naïve Bayes classifier, the classification by way of building suitable Bayesian networks are capable of handling the presence of noise in the data as well as the missing values. In addition, the classification process utilizing Bayesian networks is robust to irrelevant parameters. Obviously, Bayesian network classifiers have the same great advantages as the naïve Bayes classifier, while being free of the main disadvantage of the latter. Indeed, in each particular case, the Bayesian network classifier is designed so that the conditional independence assumptions hold automatically for the given parameters. The efficiency of Bayesian network classifiers greatly depends on the topology of the respective network. Therefore, in cases when the independence relationships between parameters can be correctly described by several different Bayesian networks, the model that has the simplest graph structure must always be preferred.

To the defense of the naïve Bayes classifier, it is very simple to implement, and has an easy way of estimating conditional probabilities. That is why, in practice, naïve Bayes classifier is often applied if the independence assumption is not far off. As we can see, there is an understandable tradeoff between the simplicity of Bayesian network reflecting the independence structure, and the precision in the independence assumptions. In practice, for the sake of significantly simplifying the algorithm, one may choose to make some assumptions as long as they are sufficiently close to the true nature of the independence relations among the variables.

Artificial neural networks can serve as an example of the Bayesian network classifier designed for a special case. Artificial neural networks can be viewed as a special case of Bayesian networks, where the nodes are partitioned into $n$ mutually-exclusive layers, and the set of arcs represented by the links from the nodes on layer $i$ to the nodes on $i + 1, 1 \leq i \leq n$. Layer 1 is usually called the input layer, while layer $n$ is known as the output layer.

4. APPLICATIONS TO BIOMEDICAL PROBLEMS

4.1. Modeling neurite regeneration and outgrows with Bayesian networks

Bayesian networks learning procedures can be effectively applied to many biomedical problems involving massive data sets. In particular, Bayesian networks can be exploited to construct comprehensive models of complex biological processes. Take for example the processes of neurite regeneration and outgrows. An advanced study of
these processes frequently involves a data collection by performing microarray experiments. Datasets arising in microarray experiments often contain noise and missing data. Bayesian networks are well-suited for handling this stochastic nature of microarray datasets. Since the observations may be entered into Bayesian network as evidence in order to correct the noisy data or fill in the missing values.

Probabilistic graphical models, in particular Bayesian networks, have been found extremely useful for extracting meaningful biological insight from the data sets arising in DNA microarray and protein array experiments. By incorporating simpler submodels, Bayesian networks models give a concise representation of complex cellular networks. Algorithms based on axiomatic principles for inferring such models from data promote a model-based methodology for analysis and exploration, which promising capabilities has been recently demonstrated by numerous applications of this methodology to gene expression data.

Bayesian Networks learning algorithm can be utilized to perform a sophisticated statistical analysis of the data on the multi-layered scale. Another attractive feature of Bayesian networks that makes them so suitable for the purpose of building neuro-regeneration model is their ability to accurately model causal relations between variables (i.e. gene and protein expression levels). This ability allows one to model interaction of various components such as genetic interactions, protein-protein interactions, protein pathways, etc.

Moreover, Bayesian networks have been proved to be particularly useful in the analysis of data sets with the number of replicates too small for conventional statistical methods to be reliable, which is a known issue in microarray data analysis. It turns out that even in such lack of experimental replication; Bayesian Networks are capable of extracting knowledge. Additionally, Bayesian networks combined with Bayesian methods can serve as an efficient instrument for avoiding the over fitting of data. It has been shown that it is not necessary to withhold some of the available data for testing and that the Bayesian approach allows us to modify the model so that all available data can be used for training [11].

Finally, Bayesian networks appears to be a very logical choice for constructing, refining and validating the model based on a multi-paradigm approach, because they allow one to utilize knowledge discovered from the analysis of the data, which are obtained on the previous level, when building a more refined model on the next level. For instance, Bayesian learning procedure can use information about the system gathered at the PC-12 cell as a prior knowledge to infer information about the system on the primary neuron level. In the framework of Bayesian networks, different characteristics of the system of neurite regeneration are treated as random variables. Random variables consist of observed attributes (for example, the expression level of a specific gene or protein in a particular experiment) and hidden attributes that are assumed by the model (for instance, the cluster allocation of a specific gene or protein).
4.2. Using Bayesian networks and data mining for genome modeling

Long before its completion in 2000, the Human Genome Project has become one of the most famous examples of a collection of massive databases. Nowadays, genomes of many various organisms have been sequenced. From the broad view of biology, the genome of an organism is all the hereditary material found in the DNA (or RNA, for some viruses). More specifically, a genome is a complete DNA sequence of a collection of chromosomes. The current human genome sequence contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers approximately 99% of the euchromatic genome and is accurate to an error rate of about an event per 100,000 bases. Remarkably, the human genome seems to encode only 20,000 – 25,000 protein-coding genes [12, 13, 14].

Even though DNA is relatively simple and well understood chemically, the human genome’s structure is extraordinarily complex and its function is poorly understood. Only less than two percent of its bases encode proteins, and the full interaction of protein-coding sequences still remains to be established. Even less is known about the function of the roughly half of the genome that consists of highly repetitive sequences or of the remaining non-coding, non-repetitive DNA.

Clearly, the task of decoding the information encoded by the human genome, as well as genomes of other sequenced species, represents a great challenge of modern science, especially in the area of data mining. In order to completely uncover the mysteries of genome, it is necessary to learn all the hidden patterns in the massive genomic data that has been accumulated. Therefore, we must utilize many advanced data mining techniques.

A two-chain spiral of DNA encodes hereditary information by a four-letter alphabet composed of the four nucleotides, namely adenine $A$, cytosine $C$, guanine $G$, and thymine $T$. It is well-known that $C – G$ and $A – T$ are the complementary pairs of bases connecting two chains of DNA. By representing a sequence of nucleotides in a human chromosome by a simple Bayesian network, specifically, the Markov chain, we establish the fundamental principles in encoding complementary bases $A$ and $T$, $C$ and $G$ by one chain in DNA chromosomes of the human genome and genomes of other investigated organisms. This indicates an opportunity to create a model of the relationships between the letter pairs, the three adjacent bases in a triplet, etc. On the basis of obtained statistical data, it is easy to come to a conclusion that there exist exact rules for forming DNA structure valid for all species.

As reported in [15, 16], it has recently been discovered that for each chain of DNA chromosomes the following complementary relationships hold:

\[
\hat{P}_A \simeq \hat{P}_T, \quad (9)
\]
\[
\hat{P}_C \simeq \hat{P}_G, \quad (10)
\]

where $\hat{P}_A, \hat{P}_C, \hat{P}_G, \hat{P}_T$ denote the frequencies of the letters $A, C, G, T$, respectively. For a given base $i = A, C, G, T$, the frequency of $i$ is calculated as follows:

\[
\hat{P}_i = \frac{n_i}{n}, \quad (11)
\]

where $n_i$ is the total number of base $A$ in the chromosome, and $n$ is the chromosomal length. These frequencies give us the estimates for the probabilities $P_i$ of the respective
nucleotides $i$, $i = A, C, G, T$; and a nucleotide sequence of a given chromosome can be modeled by a finite Markov chain $\{X_i\}_{i=1}^n$ of length $n$, with the prior probabilities known as the initial probabilities $P_A, P_C, P_G, P_T$, such that $P_A + P_C + P_G + P_T = 1$, and the conditional probabilities called the transition probabilities $P_{ij} = Pr(X_{k+1} = j \mid X_k = i)$, where $1 \leq k \leq n$, and $i, j = A, C, G, T$. Obviously, $P_A + P_C + P_G + P_T = 1$.

The frequencies $\hat{P}_{ij}$ estimating the transition probabilities $P_{ij}$ can be computed as

$$\hat{P}_{ij} = \frac{n_{ij}}{n_{iA} + n_{iC} + n_{iG} + n_{iT}}, \quad (12)$$

where $n_{ik}$ is the number of nucleotide pairs $ik$ in the data sample. There are two problems in the above approach to estimating the probabilities of transition. First, the denominator in (12) is not being fixed. Second, the estimates in (12) are biased. However, the statistical analysis [17] shows that for $i, j = A, C, G, T$, the differences below are asymptotically normal:

$$\hat{P}_{ij} - P_{ij} \sim N(0, \sigma_{ij}^2), \quad \text{as} \quad n \to \infty, \quad (13)$$

Where the standard deviation is given by

$$\sigma_{ij} = \frac{P_{ij}(1 - P_{ij})}{nP_i}. \quad (14)$$

In addition, the differences $\hat{P}_{ij} - P_{ij}$ are asymptotically independent for any given $i = A, C, G, T$. Also note that due to the extremely large chromosomal length $n$, the associated standard deviations in (14) are very small. This shows that Markov chain gives a simple and effective model of genome. In order to completely describe a very lengthy nucleotide sequence using Markov chain, for any given chromosome, it is sufficient to obtain estimates of only 15 parameters. In fact, the Markov chain model of a genome only requires estimating three out of four initial probabilities, and twelve out of sixteen transitional probabilities.

The results of frequency computations for every chromosome of the human genome are presented in Table 1. These values can be used as the estimates for the initial probabilities when constructing a corresponding Markov chain. Table 2 gives numerical results of transitional probability estimation by the frequencies of letter pairs. From Table 2 we see that for chromosome 1, all estimates of the transitional probabilities are strictly positive. Since the chain has finite number of states, this implies that the Markov chain, which approximates chromosome 1, is ergodic. In other words, for any $i$, the estimates of $n$-step transitional probabilities $\hat{P}_{ij}^n$ converge to limiting probability $\pi_j$, as $n \to \infty$, independently on $i$; and $\pi_A + \pi_C + \pi_G + \pi_T = 1$. The above convergence is geometric.

Thus, the analysis of the collected human genome data shows that the process of chromosome encoding can be modeled by an example of Bayesian network, namely, a Markov chain with four states corresponding to each nucleotide. In addition, since the length of a human chromosome is extremely large, the sequence can be given an approximate representation in form of homogeneous ergodic Markov chain.
TABLE 1. Frequencies of four nucleotides for all chromosomes of human genome

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Adenine A</th>
<th>Cytosine C</th>
<th>Guanine G</th>
<th>Thymine T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.291</td>
<td>0.209</td>
<td>0.209</td>
<td>0.292</td>
</tr>
<tr>
<td>2</td>
<td>0.299</td>
<td>0.201</td>
<td>0.201</td>
<td>0.299</td>
</tr>
<tr>
<td>3</td>
<td>0.301</td>
<td>0.198</td>
<td>0.198</td>
<td>0.302</td>
</tr>
<tr>
<td>4</td>
<td>0.309</td>
<td>0.191</td>
<td>0.191</td>
<td>0.309</td>
</tr>
<tr>
<td>5</td>
<td>0.302</td>
<td>0.197</td>
<td>0.198</td>
<td>0.303</td>
</tr>
<tr>
<td>6</td>
<td>0.302</td>
<td>0.198</td>
<td>0.198</td>
<td>0.302</td>
</tr>
<tr>
<td>7</td>
<td>0.296</td>
<td>0.204</td>
<td>0.204</td>
<td>0.297</td>
</tr>
<tr>
<td>8</td>
<td>0.299</td>
<td>0.201</td>
<td>0.201</td>
<td>0.299</td>
</tr>
<tr>
<td>9</td>
<td>0.293</td>
<td>0.207</td>
<td>0.207</td>
<td>0.293</td>
</tr>
<tr>
<td>10</td>
<td>0.292</td>
<td>0.208</td>
<td>0.208</td>
<td>0.292</td>
</tr>
<tr>
<td>11</td>
<td>0.292</td>
<td>0.208</td>
<td>0.208</td>
<td>0.292</td>
</tr>
<tr>
<td>12</td>
<td>0.296</td>
<td>0.204</td>
<td>0.204</td>
<td>0.296</td>
</tr>
<tr>
<td>13</td>
<td>0.307</td>
<td>0.193</td>
<td>0.193</td>
<td>0.308</td>
</tr>
<tr>
<td>14</td>
<td>0.294</td>
<td>0.204</td>
<td>0.205</td>
<td>0.297</td>
</tr>
<tr>
<td>15</td>
<td>0.289</td>
<td>0.211</td>
<td>0.211</td>
<td>0.289</td>
</tr>
<tr>
<td>16</td>
<td>0.275</td>
<td>0.223</td>
<td>0.224</td>
<td>0.277</td>
</tr>
<tr>
<td>17</td>
<td>0.272</td>
<td>0.228</td>
<td>0.227</td>
<td>0.273</td>
</tr>
<tr>
<td>18</td>
<td>0.301</td>
<td>0.199</td>
<td>0.199</td>
<td>0.301</td>
</tr>
<tr>
<td>19</td>
<td>0.258</td>
<td>0.242</td>
<td>0.242</td>
<td>0.259</td>
</tr>
<tr>
<td>20</td>
<td>0.278</td>
<td>0.220</td>
<td>0.221</td>
<td>0.280</td>
</tr>
<tr>
<td>21</td>
<td>0.297</td>
<td>0.204</td>
<td>0.205</td>
<td>0.294</td>
</tr>
<tr>
<td>22</td>
<td>0.261</td>
<td>0.240</td>
<td>0.240</td>
<td>0.260</td>
</tr>
<tr>
<td>X</td>
<td>0.302</td>
<td>0.197</td>
<td>0.197</td>
<td>0.303</td>
</tr>
<tr>
<td>Y</td>
<td>0.299</td>
<td>0.199</td>
<td>0.200</td>
<td>0.301</td>
</tr>
</tbody>
</table>

TABLE 2. Frequencies of different pairs of nucleotides for chromosome 1 of human genome used for estimating the transitional probabilities

<table>
<thead>
<tr>
<th>$\hat{p}_{ij}$</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.327</td>
<td>0.172</td>
<td>0.244</td>
<td>0.257</td>
</tr>
<tr>
<td>C</td>
<td>0.349</td>
<td>0.261</td>
<td>0.048</td>
<td>0.342</td>
</tr>
<tr>
<td>G</td>
<td>0.287</td>
<td>0.211</td>
<td>0.260</td>
<td>0.242</td>
</tr>
<tr>
<td>T</td>
<td>0.218</td>
<td>0.205</td>
<td>0.250</td>
<td>0.327</td>
</tr>
</tbody>
</table>

5. SUMMARY

This chapter focuses on the relationship between Bayesian networks and data mining, and the applications of these dynamic research areas to various problems in biomedicine. A general classification procedure based on Bayesian networks is described. The procedure is a generalization of the classification technique known as Naïve Bayes. The application of the Bayesian network methods to biomedical problems is illustrated by two examples. In particular, we outline the ideas behind the study of the causal relations between various components of neurite regeneration model. In addition, we give
an example of estimation of probabilities in Markov chain model derived from the human genome data. Just as Naïve Bayes, Markov chain is a special case of an elementary Bayesian network.

With the rapid development of new technologies in collecting biological data such as microarray technologies, data mining has become essential for biomedical research. Recently, the Bayesian networks techniques have been successfully applied to establish the hidden patterns in the datasets arising in microarray experiments. There is an enormous potential in combining these two fields, in particular, to tackle complex research problems in biomedicine and biology. The symbiosis of the Bayesian networks approach and efficient data mining techniques can offer a resourceful instrument for learning the underlying biological processes from massive data.

REFERENCES