Journal of King Saud University – Computer and Information Sciences (2011) 23, 79–89

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Confidence value prediction of DNA sequencing with Petri net model

Raed I. Hamed [a,*], Syed I. Ahson [b]

[a] Department of Computer Science, Faculty of Computer, University of Anbar, Al-Anbar, Iraq
[b] Patna University, Ashok Rajpath, Patna 800 005, India

Received 4 December 2010; accepted 30 April 2011
Available online 6 May 2011

KEYWORDS

Fuzzy Petri net;
Fuzzy logic;
DNA sequencing;
Modeling and simulation

Abstract In this paper, a fuzzy Petri net (FPN) approach to modeling fuzzy rule-based reasoning is proposed to determining confidence values for bases called in DNA sequencing. The proposed approach is to bring DNA bases-called within the framework of a powerful modeling tool FPN. The three input features in our fuzzy model-the height, the peakness, and the spacing of the first most likely candidate (the base called) and the peakness and height for the second likely candidate can be formulated as uncertain fuzzy tokens to determines the confidence values. The FPN components and functions are mapped from the different type of fuzzy operators of If-parts and Then-parts in fuzzy rules. The validation was achieved by comparing the results obtained with the FPN model and fuzzy logic using the MATLAB Toolbox; both methods have the same reasoning outcomes. Our experimental results suggest that the proposed models, can achieve the confidence values that matches, of available software.

© 2011 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.

1. Introduction

A major challenge of modeling biological systems is that conventional methods based on physical and chemical principles require data that is difficult to accurately and consistently obtain using either conventional biochemical or high throughput technologies, which typically yield noisy, semi-quantitative data (often in terms of a ratio rather than a physical quantity) (Fitch and Sokhansanj, 2000). Various kinds of models have been studied to express biological systems such as differential equations (Novak et al., 1998; Chen et al., 1999), Boolean networks (Liang et al., 1998; Akutsu et al., 1999), Petri Nets (Matsuno et al., 2000, 2003; Fujita et al., 2004), Bayesian networks (Husmeier, 2003) and artificial neural networks (Vohradsky, 2002). The above-mentioned papers are dedicated to the applications of different methods to genetic networks

* Corresponding author.
E-mail addresses: raed.inf@gmail.com (R.I. Hamed), drsiahson@yahoo.com (S.I. Ahson).

1319-1578 © 2011 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.

Peer review under responsibility of King Saud University.
doi:10.1016/j.jksuci.2011.05.004

Production and hosting by Elsevier

and show that these methods are suitable to model special molecular biological systems.

The explosion in the number of genomic datasets generated with tools such as high throughput DNA sequencing machines and DNA microarrays has created a critical need for resources that facilitate the interpretation of large-scale biological data. A new mathematics and novel methodologies are required to contribute to the conceptual or complex theoretical framework in which biologists study organisms. One such tool is fuzzy logic that can satisfy the need for a conceptual framework and provide a systematic and unbiased way to perform this transformation.

Since fuzzy logic system has been successfully used in various applications and large-scale complex systems exist everywhere in our society, this encourages its complex applications that have large amounts of rules and require real-time responses. Petri net theory and fuzzy logic exhibit a graphical and mathematical formalism to model, and simulate the biological systems. Fuzzy Petri net (PN) is a successful tool for describing and studying information systems. Incorporating the fuzzy logic with Fuzzy Petri nets has been widely used to deal with fuzzy knowledge representation and reasoning (Raed and Ahson, 2010; Lukas and Ralf, 2008; Chen et al., 1990; Raed and Ahson, 2010). It has also proved to be a powerful representation method for the reasoning of a rule-based system. Such an approach is appropriate for the case where a state of the modeled system corresponds to a marking of the associated FPN. The motivation for the development of the FPN model is to fuse the benefits of fuzzy logic (i.e. effectively manage uncertain or corrupted inputs, natural linguistic structure, etc.) with FPN techniques.

The advantages of using FPNs in fuzzy rule-based reasoning systems include (Chen et al., 1990; Bostan-Korpeoglu and Yazici, 2007): (1) the graphical representation of FPNs model can help to visualize the inference states and modify fuzzy rule bases; (2) the analytic capability, which can express the dynamic behavior of fuzzy rule-based reasoning. Evaluation of markings is used to simulate the dynamic behavior of the system. The explanation of how to reach conclusions is expressed through the movements of tokens in FPNs (Bostan-Korpeoglu and Yazici, 2007). The field of fuzzy Petri nets may have an important impact in understanding how biological systems work, giving at the same time a way to describe, manipulate, and analyse them. Given the complexity of systems begin studied, biologists need a modeling and simulation framework to make sense of large-scale data and intelligently design traditional bench-top experiments that provide the most biological insight. Following its first application of modeling the dynamic biological systems (Lukas and Ralf, 2008), fuzzy Petri nets as a new tool for predicting the confidence values for each base called in DNA sequencing are investigated in this paper.

The method presented in this paper develops a fuzzy Petri net model that can predict the confidence values for each base called in DNA sequencing. This approach here utilizes the information that is gathered at the base, for more information (see Ressom et al., 2005). This includes information on the height, peakness, and spacing of the base under consideration and the next likely base. In order to validate our approach, we compare our method to the fuzzy logic toolbox of MATLAB. The comparison is made in terms of the confidence value measure of the bases called in DNA sequencing. The similarity that we have discovered is that they both have the same conclusions.

The organization of this paper is as follows: in Section 2, fuzzy Petri nets are described. In Section 3, the formulation of fuzzy sets and linguistic variables are presented. In Section 4, we explain the details of methods of modeling DNA bases called together with fuzzy Petri net as a new tool for modeling DNA bases called are investigated in this paper. Section 5 describes the experimental and simulation results. Finally, we presented the conclusions of our model in Section 6.

2. Fuzzy Petri nets

2.1. Formal definition of fuzzy Petri nets

Chen et al. (1990) presented a new knowledge representation by means of fuzzy Petri nets (FPN). A fuzzy Petri net model allows a structural representation of knowledge and has got a systematic procedure for supporting the fuzzy reasoning process (Raed and Ahson, 2010). Formally, a fuzzy Petri net structure is defined as follows (Chen et al., 1990):

The tuple $FPN = (\mathrm{P, T, D, I, O, F}, \alpha, \beta)$ is called a fuzzy Petri net if:

1. $P = \{p_1, p_2, \ldots, p_n\}$ is a finite set of places, corresponding to the propositions of FPRs;
2. $T = \{\mathrm{t}_1, t_2, \ldots, t_n\}$ is a finite set of transitions, $P \cap T = \emptyset$, corresponding to the execution of FPRs;
3. $D = \{d_1, d_2, \ldots, d_n\}$ is a finite set of propositions of FPRs. $P \cap T \cap D = \emptyset, |P| = |D|$, d_i $(i = 1, 2, \ldots, n)$ denotes the proposition that interprets fuzzy linguistic variables, such as: very low, low, lnorm , eug , hnorm, as in our model;
4. I: $P \times T \rightarrow \{0, 1\}$ is an $n \times m$ input incidence matrix defining the directed arcs from propositions (P) to rules (T). $I(p_i, t_j) = 1$, if there is a directed arc from p_i to t_j, and $I(p_i, t_j) = 0$; if there is no directed arcs from p_i to t_j, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$.
5. O: $P \times T \rightarrow \{0, 1\}$ is an $n \times m$ is an output incidence matrix defining the directed arcs from rules to propositions. $O(p_i, t_j) = 1$, if there is a directed arc from t_j to p_i, and $O(p_i, t_j) = 0$; if there is no directed arcs from t_j to p_i, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$.
6. $F = \{\mu_1, \mu_2, \ldots, \mu_m\}$ where μ_i denotes the certainty factor (CF $= \mu_i$) of R_i , which indicates the reliability of the rule R_i, and $\mu_i \in [0, 1]$;
7. α: $P \rightarrow [0, 1]$ is the function which assigns a token value between zero and one to each place;
8. β: $P \rightarrow D$ is an association function, a bijective mapping from a set of places to a set of propositions.

Moreover, this model can be enhanced by including a function $Th : T \rightarrow [0, 1]$ which assigns a threshold value $Th(t_j) = \lambda_j \in [0, 1]$ to each transition t_j, where $j = 1, \ldots, m$. Further more, a transition is enabled and can be fired in FPN models when values of tokens in all input places of the transition are greater than its threshold.

A token value in place $p_i \in P$ is denoted by $\alpha(p_i) \in [0, 1]$, $\alpha(p_i) = y_i$, $y_i \in [0, 1]$ and $\beta(p_i) = d_i$. This states that the degree of the truth of proposition d_i is y_i. A transition t_i is enabled if $\forall p_i \in I(t_i)$, $y_i > 0$. If this transition t_i is fired, tokens are removed from input places $I(t_i)$ and a token is deposited onto each of the output places $O(t_i)$. This token's membership value to the place p_k, (i.e. $y_k = \alpha(p_k)$), is part of the token and gets calculated within the transition function. It is easy to see that $CF \in [0, 1]$. If $CF = 1$ then we will say that a given rule is deterministic. Otherwise (i.e., if $CF < 1$), we will say that the given rule is non-deterministic. In a fuzzy Petri net, different types of rules can be represented. The general ones are:

1. A simple fuzzy production rule:
   IF d_i *THEN* d_k ($CF_j = f(t_j)$);
2. A composite conjunctive rule:
   IF d_1 *AND* d_2 *AND* ... *AND* d_j *THEN* d_k ($CF_j = f(t_j)$);
3. A composite disjunctive rule:
   IF d_1 *OR* d_2 *OR* ... *OR* d_j *THEN* d_k ($CF_j = f(t_j)$);

Type 1. A simple fuzzy production rule. $I(t_J) = \{p_i\}$, $O(t_j) = \{p_i\}$, $f(t_j) = CF$, $\beta(p_i) = d_i$ and $\alpha(p_i) > 0$, where $1 \leqslant i$, $k \geqslant n$ and $1 \leqslant j \geqslant m$. It means that the degree of truth of the proposition $\beta(p_i) = d_i$ in this place p_i is equal $\alpha(p_i)$. Moreover, if the threshold value for the transition t_j is given, the transition t_j can be fired if only $\alpha(p_i) > \lambda_j$ (in otherwise this transition can always be fired). After firing the transition t_j, $\alpha(p_k) = \alpha(p_i) \times CF$.

Type 2. A composite conjunctive rule. $I(t_j) = P = \{p_1, p_2, \ldots, p_l\}$, $O(t_j) = p_k$, $f(t_j) = CF$, $\beta(P) = [d_1, d_2, \ldots, d_l]$ and $\alpha(P) = [\alpha(p_1), \alpha(p_2), \ldots, \alpha(p_n)]$, where $1 \leqslant l$, $k \geqslant n$ and $1 \leqslant j \geqslant m$. After firing the transition$t_j : \alpha(p_k) = \min_{1 \leqslant i \leqslant l} \alpha(p_i) \times CF$.

Type 3. A composite disjunctive rule $I(t_j) = P = \{p_1, p_2, \ldots, p_l\}$, $O(t_j) = p_k$, $f(t_j) = CF$, $\beta(P) = [d_1, d_2, \ldots, d_l]$ and $\alpha(P) = [\alpha(p_1), \alpha(p_2), \ldots, \alpha(p_n)]$, where $1 \leqslant l$, $k \geqslant n$ and $1 \leqslant j \geqslant m$. After firing the transition$t_j : \alpha(p_k) = \max_{1 \leqslant i \leqslant l} \alpha(p_i) \times CF$.

As an example of reasoning rules, we give the Type-2 and Type-3 models in Figs. 1 and 2, respectively.

The certainty factor ($CF = \mu_i$) value is designed to reflect the way the experts think (Qu and Shirai, 2003). In this article, assuming that the belief strength of a fuzzy rule assigned by human expert is $CF = 1$, which represents that the fuzzy rule is completely believable, and denoted by a composite conjunctive rule.

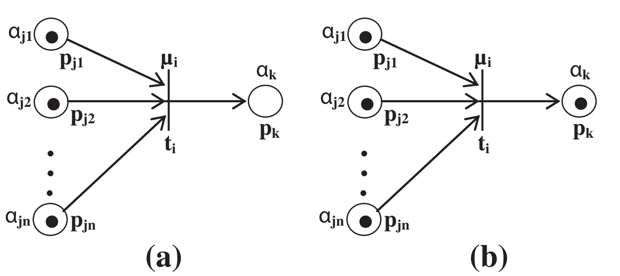

Figure 1 Tapy-2 fussy rule representation inFRPNs. (a) Before firing transitions and (b) after firing transitions.

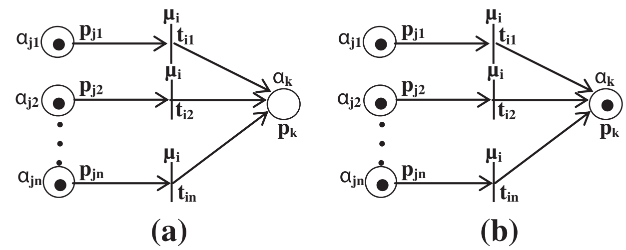

Figure 2 Tapy-3 fussy rule representation in FRPNs. (a) Before firing transitions, (b) after firing transitions.

2.2. Fuzzy Petri net model

The Mamdani fuzzy inference system (Mamdani and Assilian, 1975) was proposed as the first attempt to control a steam engine and boiler combination by a set of linguistic control rules obtained from experienced human operators. The output membership functions of Mamdani models are fuzzy sets, which can incorporate linguistic information into the model. The computational approach described in this paper is Mamdani fuzzy Petri net (MFPN) that is able to overcome the drawbacks specific to pure Petri nets.

Fuzzy models describing dynamic processes compute the states $x(t+1)$, at a time instant $t+1$, from the information of the inputs $x(t)$ and $u(t)$, at time instant t:

$$x(t+1) = f(x(t), u(t)), \tag{1}$$

where $f(\cdot)$ is a fuzzy model with the structure shown in Fig. 3. In the input Layer (***Layer* 1**), as shown in the next equation no calculation is done in this layer. Each node, which corresponds to the inputs, $x(t)$ and $u(t)$, only transmits input value to the next layer directly. The certainty factor of the transitions in this layer is unity.

$$O^{(1)} = x(t), u(t), \tag{2}$$

where $x(t)$ and $u(t)$ are the expression value of the ith gene at time instant t, and $O_i^{(1)}$ is the ith output of layer 1. Nodes in ***layer* 2** are called input term nodes. Where the values of the inputs, $x(t)$ and $u(t)$, and the outputs, $x(t+1)$, can be assigned linguistic labels, e.g., 'low-expressed' (L), 'medium-expressed' (M), and 'high-expressed ' (H). The output link of layer 2, represented as the membership value, specifies the degree to which the input value belongs to the respective label. Linguistic rules can be formulated that connect the linguistic labels for $x(t)$ and $u(t)$ via an IF-part, called an antecedent of a rule and the THEN-part, also called a consequent of the rule which determines the resulting linguistic label for $x(t+1)$. The structure of a single rule can thus be presented as follows:

IF ($x(t)$ is $A_{x(t)}$) AND ($u(t)$ is $A_{u(t)}$), THEN ($x(t+1)$ is $A_{x(k+1)}$), where $A_{x(t)}$, $A_{u(t)}$, and $A_{x(t+1)}$ are the linguistic labels for $x(t)$, $u(t)$, and $x(t+1)$, respectively, generated for the data points. The antecedent defines the condition, and the consequent – the conclusion which will be implemented if the condition is true. The antecedent membership functions are the membership functions appearing in the IF-part of the rule in layer 2 and the consequent membership functions are the membership functions appearing in the THEN-part in layer 4. As shown in the following equation, membership functions are represented as Trapezoidal form.

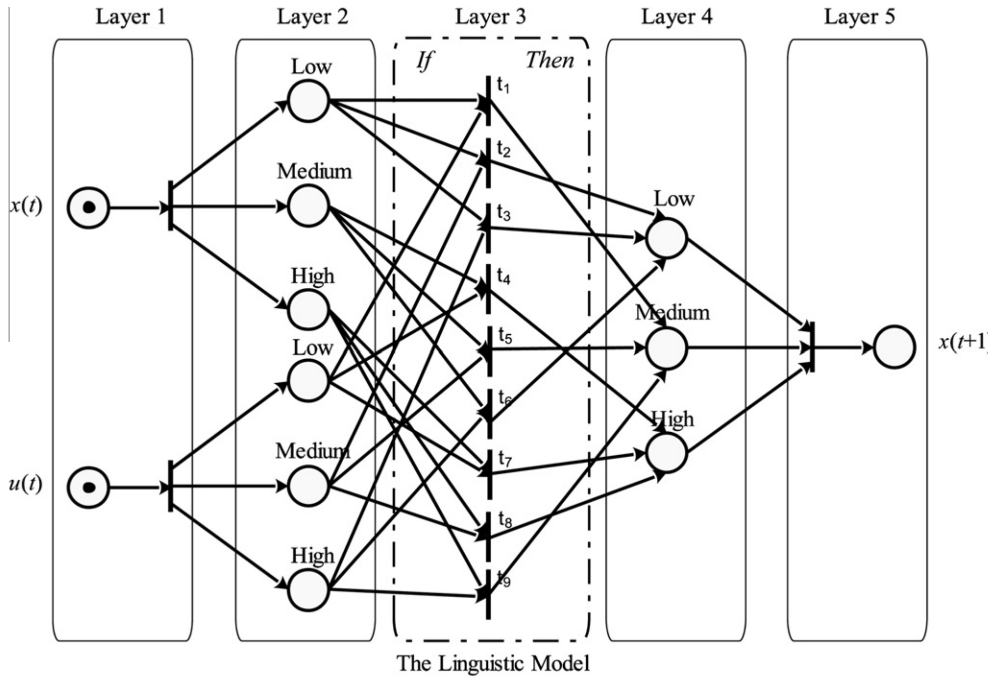

Figure 3 Fuzzy layer structure of mamdani fuzzy Petri net with $x(t)$, $u(t)$ inputs and $x(t+1)$ output.

$$O^{(2)} = \mu_A(x) = \begin{cases} 0, & x \le a, \\ (x-a)/(b-a), & a \le x \le b, \\ 1, & b \le x \le c, \\ (d-x)/(d-c), & c \le x \le d, \\ 0, & d \le x, \end{cases} \quad (3)$$

where μ_A refers to the degree to which $x$belongs to the linguistic label A and the parameters $\{a, b, c, d\}$ (with $a < b < c < d$) determine the x coordinates of the four corners of the underlying trapezoidal membership function.

Nodes in ***layer* 3** are called rule based nodes. A node in this layer combines the antecedent part of a fuzzy rule using a T-norm operation. We use the AND operation on each rule node as minimum operation. The output of each node represents the firing strength of the corresponding fuzzy rule.

Nodes in ***layer* 4** are called output term nodes and this layer is called the consequent layer. Each output term node represents a fuzzy set (described by a Trapezoidal function) obtained by fuzzy Petri net structure. Different nodes in layer 3 may be connected to a same node in this layer, meaning that the same consequent is specified for different rules. The function of each output term node performs the following fuzzy OR operation:

$$O^{(4)} = \sum_i t_i^{(4)}. \quad (4)$$

In the following equation, the symbol $t_i^{(4)}$ denotes the ith input of a node in the 4th layer. To integrate the fired rules which have the same consequent part. The above fuzzy OR operation is a modified bounded sum operation in fuzzy theory (Raed and Ahson, 2010).

Each node in ***layer* 5** is called an output linguistic node and corresponds to one output linguistic variable. This layer performs the defuzzification operation. The nodes in this layer together with the links attached to them accomplish this task. We need to find the crisp output of layer 5 by finding the "center of gravity" method as the defuzzification using the output of the node in layer 4:

$$O^{(5)} = \frac{\sum_i O_i^{(4)} y_i}{\sum_i O_i^{(4)}}, \quad (5)$$

where the symbol $O_i^{(4)}$ denotes the node output in layer 4, and y_i, is the center of the membership function of the term of the output linguistic variable. We apply the center-of-gravity method because the aggregate implication results in a new fuzzy output set, while in fact we need a single crisp output. Applying the *maximum*operation to all the resulting implications performs the aggregation. The linguistic terms of input and output nodes and the centers of the membership functions of linguistic terms (Trapezoidal membership functions employed here) should be correctly determined in order for the fuzzy inference system to produce corresponding outputs according to inputs in training data.

3. Formulation of fuzzy sets and linguistic variables

Using the *IF-THEN* statements, the estimation procedure for the FPN model can be described by the following two step process. (1) Evaluate the antecedent proposition: fuzzy membership functions are used to determine the extent by which each antecedent (IF) "fires". (2) Evaluate the consequent proposition: the "fired" consequents (THEN) are aggregated into predictions for the outputs. Before the above steps can be discussed in detail, the fuzzy membership function is discussed.

The input variables of the antecedent proposition considered for the fuzzy rule include the *height* (H_{called} and H_{2nd}), *peakness* (P_{called} and P_{2nd}), and *spacing* ($|\Delta S_{\text{previous}}|$ and |

$\Delta S_{next}|$) for more information (see Ressom et al., 2005). The output variables of the fuzzy reasoning include the confidence value of DNA bases called. In reference to Ressom et al. (2005), the fuzzy linguistic terms and trapezoidal membership function for our fuzzy Petri net models are included in the fuzzy rule and illustrated in Fig. 4.

Any input value can be described through a combination of membership values in the linguistic fuzzy sets. Note that incorporation of linguistic fuzzy sets provides a tool for natural computing (Zadeh, 1975; Zadeh et al., 2002) as the resulting system is capable of reasoning like human. In order to measure these input and output metadata universally, we normalize them into the same standard scale of $[0, 1]$. The values of linguistic variables are fuzzified to obtain the membership degree by membership function. For example, $\mu_{low}\_P_{called} = (0.35) = 0.5$, $\mu_{medium}\_P_{called} = (0.35) = 0.5$, means the value, 0.35 belongs to medium with confidence value (i.e. truth degree) of 50% while 50% belongs to low. That is, a 3-d membership vector for the fuzzy sets low, medium, and high corresponding to fuzzy peaknesses (P_{called}) is generated and is given by:

$$VP_{called} = [\mu_{flat}\_P_{called}, \mu_{medium}\_P_{called}, \mu_{sharp}\_P_{called}]^T. \quad (6)$$

Similarly, peaknesses (P_{2nd}), height (H_{called}), height (H_{2nd}), spacing (ΔS_{next}), and spacing ($\Delta S_{previous}$) are defined as:

$$VP_{2nd} = \left[\mu_{flat}\_P_{2nd}, \mu_{medium}\_P_{2nd}, \mu_{sharp}\_P_{2nd}\right]^T,$$

$$VH_{called} = \left[\mu_{vlow}\_H_{called}, \mu_{low}\_H_{called}, \mu_{medium}\_H_{called}, \mu_{high}\_H_{called}, \mu_{vhigh}\_H_{called}\right]^T,$$

$$VH_{2nd} = \left[\mu_{vlow}\_H_{2nd}, \mu_{low}\_H_{2nd}, \mu_{medium}\_H_{2nd}, \mu_{high}\_H_{2nd}, \mu_{vhigh}\_H_{2nd}\right]^T,$$

$$V\Delta S_{next} = \left[\mu_{small}\_\Delta S_{next}, \mu_{medium}\_\Delta S_{next}, \mu_{large}\_\Delta S_{next}\right]^T,$$

$$V\Delta S_{previous} = \left[\mu_{small}\_\Delta S_{previous}, \mu_{medium}\_\Delta S_{previous}, \mu_{large}\_\Delta S_{previous}\right]^T.$$

4. Methods and modeling of DNA bases called

4.1. Methods

DNA sequencing analysis software, commonly known as base-callers, process the raw data of the trace and estimates the most likely set of bases in the original DNA sample (Berno, 1996). In principle, the DNA sequence can be obtained from the electropherogram by associating each dominant peak with the corresponding base type and by preserving the order of the peaks. The term DNA sequencing usually refers to the process of determining the ordered sequence of bases which constitutes the genetic code. Four different bases (nucleotide types) are present in a DNA strand: *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T) (Genome Project Information, xxxx). The most widely used base-caller today is Phred program (Ewing et al., 1998). *Phred* uses a four-phase procedure to determine the sequence of bases in a segment of DNA. The *Phred* program is used to call the bases in the raw trace files generated by automated sequencing machines (more information about *Phred* see Ewing et al., 1998).

The examine trace features used in *Phred* base calling system include the following:

1. *Peak spacing*: this feature is the ratio of the largest peak-to-peak ratio to the smallest peak-to-peak ratio within a window of seven peaks.
2. *Uncalled/called ratio*: this feature is the ratio of the amplitudes of the largest uncalled peak to the smallest called peak.
3. *Uncalled/called ratio2*: the only difference between the second and third features is the window size used, which was seven and three peaks respectively.
4. *Peak resolution*: this final feature is the number of bases between the current base and the next unresolved base (in the phred system an unresolved base call is labeled with an N).

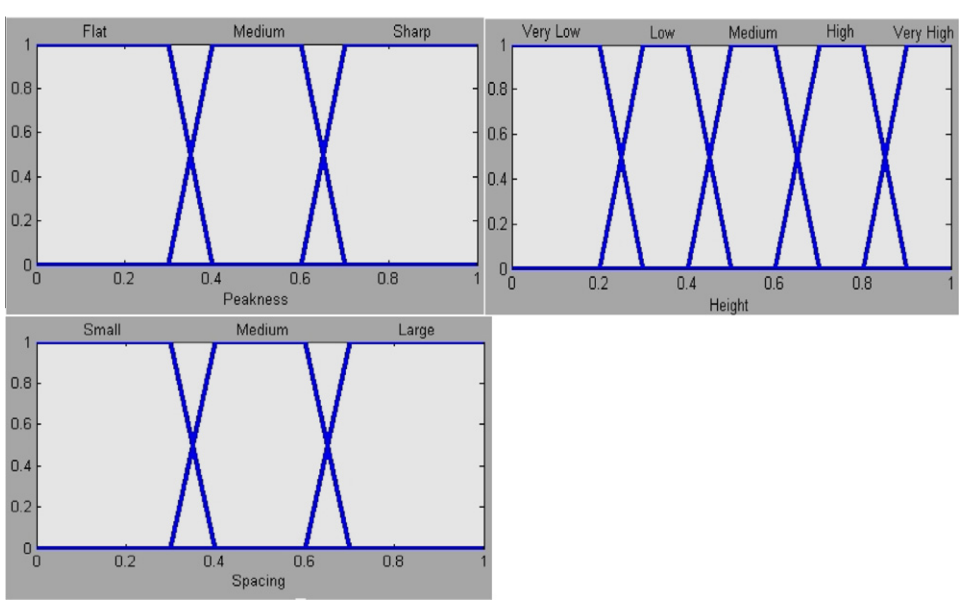

Figure 4 Trapezoidal membership functions of the fuzzy variables, "peakness", "height", and "spacing".

Many biological processes and objects are intrinsically fuzzy as their properties and behaviors contain uncertainty information. The applications of fuzzy logic are ideal to describe some biological processes/objects and provide good tools for many bioinformatics problems. Recently, there has been effort to address base-calling in sound modeling framework (Ressom et al., 2005) in order to improve the accuracy and base confidence value estimation. Ressom et al. (2005) develop a fuzzy logic algorithm that can predict the confidence values for each base called in DNA sequencing. However, this technique utilizes the information that is gathered at the base. This includes information on the *height, peakness*, and *spacing* of the base under consideration and the next likely base. In particular we refer to the DNA bases called process known as fuzzy logic method. Following its first application of fuzzy logic to develop confidence measure for assessing the accuracy of DNA bases called (Ressom et al., 2005), fuzzy Petri net as a new tool for modeling DNA bases called are investigated in this paper.

4.2. Fuzzy Petri net approach to modeling for DNA bases called

Fig. 5 shows an overview of the fuzzy modeling process. This schematic indicates the components that have to be defined for our application, including three inputs considered in the model *Peakness*, *Height*, and *Spacing* by which the fuzzy *Confidence* value can be inferred in terms of five degrees very low (VL), low (L), medium (M), high (H), and very high (VH).

Procedure FPN presented below to predict the confidence value of the three fuzzy sub-systems (i.e. fuzzy Peakness, fuzzy Height, and fuzzy Spacing). When implemented, a FPN, can be used as an inference engine of the main fuzzy system comprises of four models, namely, (i) the FPN peakness modle, (ii) the FPN height module, (iii) the FPN spacing model, and (iv) the FPN main system model. After the construction of the FPN model is over, we initialize the beliefs (truth degree) of the propositions/predicates mapped at the appropriate places. The implementation of a FPN is realized by firing transitions (rules). A transition t fires instantly as soon as it is enabled. Transitions (T) is enabled if all its input places have tokens whose truth degrees are greater than or equal their thresholds. For the antecedent-consequent pairs of each rule, a transition is created and the connectivity between the antecedent parts and the transition, and connectivity between the transition and the consequent parts are established following the rules. After a rule is encoded as part of the fuzzy Petri net, it is discarded from the set of rules. The process described above is repeated for each rule. The procedure terminates when all the rules have been encoded into corresponding place-transition-arc triplets. However the degrees of truth of antecedent parts are given, we want to compute the degrees of truth of consequent parts. Based on the concepts of FPN we presented the following procedure for computation of the subsystems of the Fig. 6(a)–(c).

STEP 1: Enter the required variables (Peakness, Height, and Spacing) for the fuzzy rule in the model.
STEP 2: According to the trapezoidal formula calculate the membership degree of the proposition of variables.
STEP 3: Calculate the firing strength by the composition AND operator (MIN).
STEP 4: Calculate the maximum firing strength by the composition OR operator (MAX).

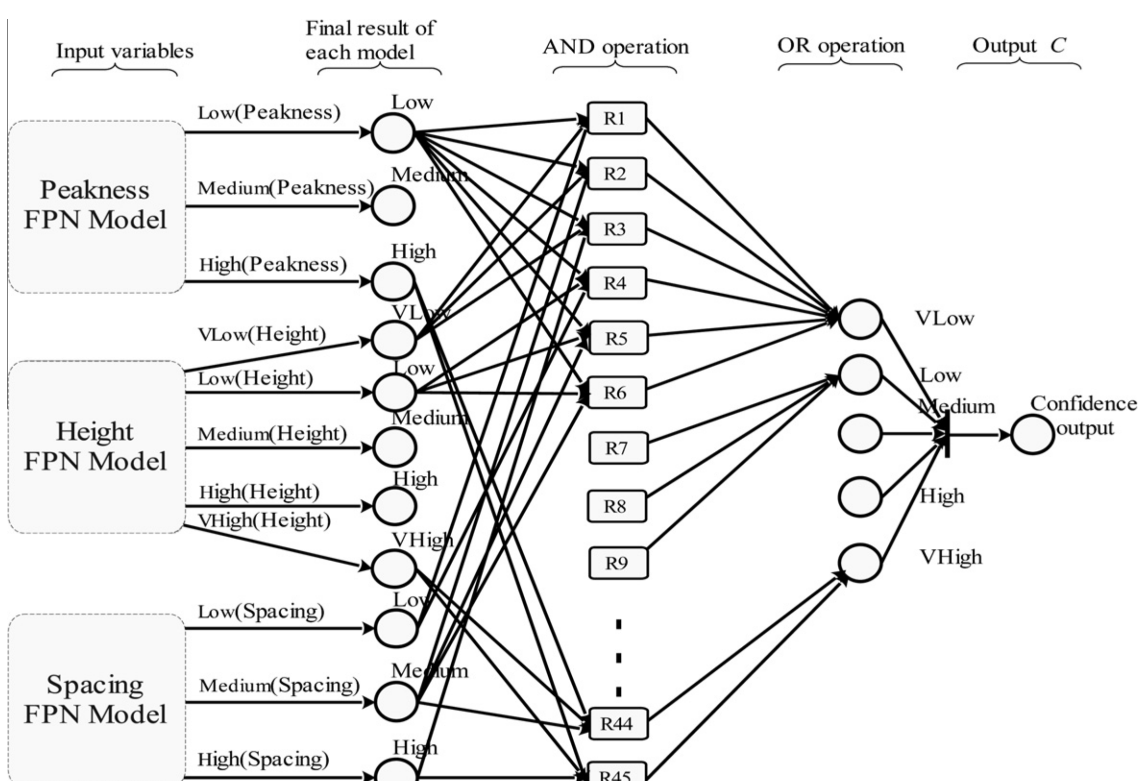

Figure 5 Block diagram of fuzzy inference reasoning structure.

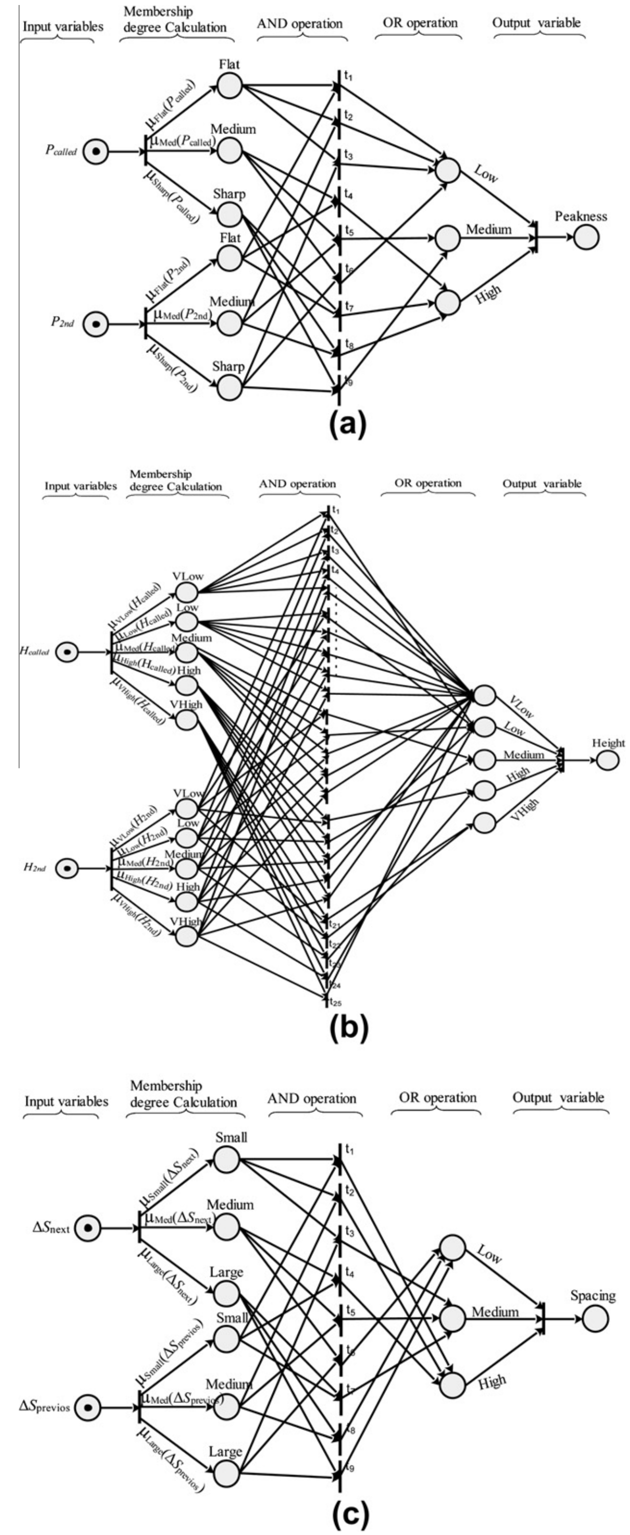

Figure 6 Instances of modeling fuzzy production rules of (a) a peaknees subsystem, (b) a height subsystem, and (c) a spacing subsystem – by a fuzzy Petri net.

STEP 5: Calculate a conclusion of the output for each subsystem by defuzzyfication formula.

In this paper, the implication relationship of the antecedent and consequent proposition in fuzzy Petri net models is used to establish the elements for fuzzy rules follow the Mamdani model (Mamdani and Assilian, 1975). As we have show in Fig. 5 it is possible to compute the firing composition as follows:

$R1$ = Low(Peakness) AND VLow(Height) AND Low(Spacing) Then VLow(Confidence),
$R2$ = Low(Peakness) AND VLow(Height) AND Medium(Spacing) Then VLow(Confidence),
$R3$ = Low(Peakness) AND VLow(Height) AND High(Spacing) Then VLow(Confidence),
⋮
$R44$ = High(Peakness) AND VHigh(Height) AND Medium(Spacing) Then VHigh(Confidence),
$R45$ = High(Peakness) AND VHigh(Height) AND High(Spacing) Then VHigh(Confidence),

4.3. Fuzzy Petri nets for variables modeling

As illustrated in Fig. 6, the models of Peakness, Height, and Spacing are used to describe the fuzzy inference reasoning system. Fig. 6(a) shows the contents of the fuzzy production rules in the peakness model. Fig. 6(b) shows of the contents of the fuzzy production rules in the height model. Fig. 6(c) shows the contents of the fuzzy production rules in the spacing model by means of the FPNs technique. The properties of the proposition set of places and the firing transitions for peaknees, height, and spacing are described as follows:

(1) P_{called}, P_{2nd}, H_{called}, H_{2nd}, ΔS_{next} and $\Delta S_{\text{previous}}$: represents the ‘‘input data’’. Each place has an input variable.
(2) The first transition: represents the ‘‘membership function’’ transition of input information. It represents the transition distribution of linguistic variables for each antecedent proposition of a rule.
(3) Flat, Medium, Sharp, VLow, Low, High, VHigh, Small, and Large: represents the ‘‘membership degree’’ place of input information.
(4) $T_{peakness} = (t_1, t_2, \ldots, t_9)$, $T_{height} = (t_1, t_2, \ldots, t_{24})$, and $T_{spasing} = (t_1, t_2, \ldots, t_9)$: represents the ‘‘firing strength’’ transition of the activated fuzzy rules. It is calculated by the composition operator MIN.
(5) VLow, Low, High, Medium, and VHigh: represents the value of a ‘‘consequent proposition’’ place. Each place with a token represents the value of a consequent proposition of a fuzzy rule. It is calculated by the composition operator MAX of the activated rules with the highest firing strength.
(6) The value of a consequent proposition of the winning fuzzy rule from the available rules in each model is calculated by the centroid of the aggregate.
(7) Peakness, Height, and Spacing: is a ‘‘final decision-making’’ place. Each place with a token represents the final result from each subsystem.

Transitions for each rule in the models represent the firing strength of the rule, uses the fuzzy operator ‘‘*AND*’’ or ‘‘*OR*’’ to perform MIN or MAX composition operation. For example, with the peakness model we have the following MIN rules:

$R1(t_1) = \text{MIN}(\mu_{\text{Flat}}\ (P_{\text{called}}),\ \mu_{\text{Flat}}\ (P_{\text{2nd}}))$.
$R2(t_2) = \text{MIN}(\mu_{\text{Flat}}\ (P_{\text{called}}),\ \mu_{\text{Med}}\ (P_{\text{2nd}}))$.
$R3(t_3) = \text{MIN}(\mu_{\text{Flat}}\ (P_{\text{called}}),\ \mu_{\text{Sharp}}\ (P_{\text{2nd}}))$.

⋮

$R9(t_9) = \text{MIN}(\mu_{\text{Sharp}}\ (P_{\text{called}}),\ \mu_{\text{Sharp}}\ (P_{\text{2nd}}))$.

The reasoning steps for each model are described as follows.

Algorithm 1. The reasoning algorithm for confidence value prediction

STEP 1: The knowledge base is described by rules for problem of the confidence value prediction for bases called in DNA sequencing.
STEP 2: The fuzzy Petri net of rules in the knowledge base is modeled.
STEP 3: Enter the required variables (P_{called} and P_{2nd} of Peakness model, H_{called} and H_{2nd} of Height model, and ΔS_{next} and $\Delta S_{\text{previous}}$ of ΔSpacing model) for each fuzzy Petri net model.
STEP 4: According to the Trapezoidal formula calculate the membership degree of the proposition of variables.
STEP 5: Calculate the firing strength by the composition AND operator (MIN).
STEP 6: Calculate the maximum firing strength by the composition OR operator (MAX).
STEP 7: Calculate a conclusion of the output for each sub-system by defuzzyfication formula.
STEP 8: Then the value of the defuzzyfication of Peakness, Height, and ΔSpacing models corresponding to the confidence value.

5. Experimental and simulation results

As shown in Fig. 6(a)–(c), the fuzzy rule base of the reasoning process as a part in the main system to determine the confidence value is constructed of three models. Here we describe existing variables (peakness, height, and spacing) following the method (Ressom et al., 2005) which has been used for comparative analysis. The fuzzy membership functions of these input variables are described in Fig. 2. We input a crisp data (i.e. P_{called}, P_{2nd}, H_{called}, H_{2nd}, ΔS_{next} and $\Delta S_{\text{previous}}$) into those corresponding membership functions, and get the membership degree for all variables as listed in the fourth column of Table 1. To explain our method a part of a DNA sequence that involves six bases (ATCTCG) is presented as listed in the third column of Table 1. Table 1 shows the P_{called}, P_{2nd}, H_{called}, H_{2nd}, ΔS_{next} and $\Delta S_{\text{previous}}$ for the six bases. For example, for the base G the normalized value for each input data as follows: $P_{\text{called}} = 1$, $P_{\text{2nd}} = 0.62$; $H_{\text{called}} = 0.98$, $H_{\text{2nd}} = 0.49$; $\Delta S_{\text{next}} = 0.28$, and $\Delta S_{\text{previous}} = 0.3$. The membership degrees of these input data are calculated by Trapezoidal membership functions. These membership function value can be used as the truth degree of each antecedent proposition in our FPN models. For example, with base G the truth degree of the proposition listed as:

Peakness base G	*Height base G*	*Spasing base G*
$\mu_{\text{Flat}}(P_{\text{called}}) = 0$	$\mu_{\text{VLow}}(H_{\text{called}}) = 0$	$\mu_{\text{Small}}(\Delta S_{\text{next}}) = 1$
$\mu_{\text{Medium}}(P_{\text{called}}) = 0$	$\mu_{\text{Low}}(H_{\text{called}}) = 0$	$\mu_{\text{Medium}}(\Delta S_{\text{next}}) = 0$
$\mu_{\text{Sharp}}(P_{\text{called}}) = 1$	$\mu_{\text{Medium}}(H_{\text{called}}) = 0$	$\mu_{\text{Large}}(\Delta S_{\text{next}}) = 0$
$\mu_{\text{Flat}}(P_{\text{2nd}}) = 0$	$\mu_{\text{High}}(H_{\text{called}}) = 0$	$\mu_{\text{Small}}(\Delta S_{\text{Previous}}) = 1$
$\mu_{\text{Medium}}(P_{\text{2nd}}) = 0.8$	$\mu_{\text{VHigh}}(H_{\text{called}}) = 1$	$\mu_{\text{Medium}}(\Delta S_{\text{Previous}}) = 0$
$\mu_{\text{sharp}}(P_{\text{2nd}}) = 0.2$	$\mu_{\text{VLow}}(H_{\text{2nd}}) = 0$	$\mu_{\text{Large}}(\Delta S_{\text{Previous}}) = 0$
	$\mu_{\text{Low}}(H_{\text{2nd}}) = 0.1$	
	$\mu_{\text{Medium}}(H_{\text{2nd}}) = 0.9$	
	$\mu_{\text{High}}(H_{\text{2nd}}) = 0$	
	$\mu_{\text{VHigh}}(H_{\text{2nd}}) = 1$	

For each input data the firing strength of each activated rule is calculated by the MIN and MAX composition operator, respectively. It yields

Peakness base G

$FR_1: \text{MIN}(0, 0) = 0,$
$FR_2: \text{MIN}(0, 0.8) = 0,$
$FR_3: \text{MIN}(0, 0.2) = 0,$
$FR_4: \text{MIN}(0, 0) = 0,$
$FR_5: \text{MIN}(0, 0.8) = 0,$
$FR_6: \text{MIN}(0, 0.2) = 0,$
$FR_7: \text{MIN}(1, 0) = 0,$
$FR_8: \text{MIN}(1, 0.8) = 0.8,$
$FR_9: \text{MIN}(1, 0.2) = 0.2,$

Low: $\text{MAX}(FR_1, FR_2, FR_3, FR_6) = \text{MAX}(0,0,0,0) = 0$, Medium: $\text{MAX}(FR_5, FR_9) = \text{MAX}(0, 0.2) = 0.2$, High: $\text{MAX}(FR_4, FR_7, FR_8\ ) = \text{MAX}(0,0,0.8) = 0.8$.

Height base G

$FR_1: \text{MIN}(0, 0) = 0,$	$FR_{14}: \text{MIN}(0, 0) = 0,$
$FR_2: \text{MIN}(0, 0.1) = 0,$	$FR_{15}: \text{MIN}(0, 0) = 0,$
$FR_3: \text{MIN}(0, 0.9) = 0,$	$FR_{17}: \text{MIN}(0, 0.1) = 0,$
$FR_4: \text{MIN}(0, 0) = 0,$	$FR_{18}: \text{MIN}(0, 0.9) = 0,$
$FR_5: \text{MIN}(0, 0) = 0,$	$FR_{19}: \text{MIN}(0, 0) = 0,$
$FR_6: \text{MIN}(0, 0) = 0,$	$FR_{19}: \text{MIN}(0, 0) = 0,$
$FR_7: \text{MIN}(0, 0.1) = 0,$	$FR_{21}: \text{MIN}(1, 0) = 0,$
$FR_8: \text{MIN}(0, 0.9) = 0,$	$FR_{22}: \text{MIN}(1, 0.1) = 0.1,$
$FR_9: \text{MIN}(0, 0) = 0,$	$FR_{23}: \text{MIN}(1, 0.9) = 0.9,$
$FR_{10}: \text{MIN}(0, 0) = 0,$	$FR_{24}: \text{MIN}(1, 0) = 0,$
$FR_{11}: \text{MIN}(0, 0) = 0,$	$FR_{25}: \text{MIN}(1, 0) = 0,$
$FR_{12}: \text{MIN}(0, 0.1) = 0,$	
$FR_{13}: \text{MIN}(0, 0.9) = 0,$	

VLow: MAX(FR2, FR3, FR4, FR5, FR7, FR8, FR9, FR10, FR13, FR14, FR15, FR19, FR20, FR25) = MAX(0,0,0,0,0,0,0,0,0,0,0,0,0,0) = 0,

Low: $\text{MAX}(FR_1, FR_6, FR_{12}, FR_{18}, FR_{24}) = \text{MAX}(0,0,0,0,0) = 0$, Medium: $\text{MAX}(FR_{11}, FR_{17}) = \text{MAX}(0,0) = 0$, High: $\text{MAX}(FR_{16}, FR_{23}) = \text{MAX}(0, 0.9) = 0.9$ VHigh: $\text{MAX}(FR_{21}, FR_{22}) = \text{MAX}(0, 0.1) = 0.1$

Spacing base G

$FR_1: \text{MIN}(1, 1) = 1,$
$FR_2: \text{MIN}(1, 0) = 0,$
$FR_3: \text{MIN}(1, 0) = 0,$
$FR_4: \text{MIN}(0, 1) = 0,$
$FR_5: \text{MIN}(0, 0) = 0,$
$FR_6: \text{MIN}(0, 0) = 0,$
$FR_7: \text{MIN}(1, 1) = 0,$
$FR_8: \text{MIN}(1, 0) = 0.8,$
$FR_9: \text{MIN}(1, 0) = 0.2,$

Low: $\text{MAX}(FR_6, FR_8, FR_9) = \text{MAX}(0,0,0) = 0$, Medium: $\text{MAX}(FR_3, FR_5, FR_7) = \text{MAX}(0,0,0) = 0$, High: $\text{MAX}(FR_1, FR_2, FR_4) = \text{MAX}(1,0,0) = 1$,

According to the result of *max* composition operation the defuzzification of output is used to make a final decision. We

Table 1 Membership function degree of the bases value.

Variables	Input data	Bases						Membership function value				
		A	**T**	**C**	**T**	**C**	**G**	Flat	Medium		Sharp	
Peakness	P_{called}	1.0	1.0	0.8	1.0	0.93	1.0	**A**. 0.0	0.0		1	
								T. 0.0	0.0		1	
								C. 0.0	0.0		1	
								T. 0.0	0.0		1	
								C. 0.0	0.0		1	
								G. 0.0	0.0		1	
	P_{2nd}	0.37	0.48	0.84	0.72	0.68	0.62	**A**. 0.3	0.7		0.0	
								T. 0.0	1		0.0	
								C. 0.0	0.0		1	
								T. 0.0	0.0		1	
								C. 0.0	0.2		0.8	
								G. 0.0	0.8		0.2	
								VLow	Low	Med	High	VHigh
Height	H_{called}	0.9	1.0	0.63	0.98	0.7	0.98	**A**. 0.0	0.0	0.0	0.0	1
								T. 0.0	0.0	0.0	0.0	1
								C. 0.0	0.0	0.7	0.3	0.0
								T. 0.0	0.0	0.0	0.0	1
								C. 0.0	0.0	0.0	1	0.0
								G. 0.0	0.0	0.0	0.0	1
	H_{2nd}	0.58	0.4	0.6	0.6	0.53	0.49	**A**. 0.0	0.0	1	0.0	0.0
								T. 0.0	1	0.0	0.0	0.0
								C. 0.0	0.0	1	0.0	0.0
								T. 0.0	0.0	1	0.0	0.0
								C. 0.0	0.0	1	0.0	0.0
								G. 0.0	0.1	0.9	0.0	0.0
								Small		Med	Large	
Spacing	ΔS_{next}	0.3	0.3	0.28	0.28	0.3	0.28	**A**. 1		0.0	0.0	
								T. 1		0.0	0.0	
								C. 1		0.0	0.0	
								T. 1		0.0	0.0	
								C. 1		0.0	0.0	
								G. 1		0.0	0.0	
	$\Delta S_{previous}$	0.3	0.3	0.3	0.28	0.28	0.3	**A**. 1		0.0	0.0	
								T. 1		0.0	0.0	
								C. 1		0.0	0.0	
								T. 1		0.0	0.0	
								C. 1		0.0	0.0	
								G. 1		0.0	0.0	

adopt the "center of gravity" method in Negnevitsky (2002) to solve this problem. Then, the defuzzification of peakness, height, and spacing is calculated as Peakness = 0.76, Height = 0.76, and Spacing = 0.82 by the centroid of the aggregate output membership function in the each FPNs model. Following the steps of the reasoning process, the final winning rule in Peakness FPN model is FR_8 (*IF* P_{called} *is Sharp and* P_{2nd} *is Medium THEN the Peakness is High*), which indicates that the "Peakness is High", in the Height FPN Model the final winning rule is FR_{23} (*IF* H_{called} *is VeryHigh and* H_{2nd} *is Medium THEN the Height is High*), which indicates the "Height is High", and in the Spacing FPN Model the final winning rule is FR_1 (*IF*ΔS_{next}*is Small and* $\Delta S_{previous}$*is Small THEN the Spacing is High*), which indicates the "Spacing is High".

These peakness, height, and spacing values are then imported to the antecedent propositions in the main system model to determine the confidence value Fig. 5. The fuzzy rules of main system are aggregated and defuzzied to have a crisp value of confidence value = 0.75. By calculating the centroid, which indicates the rule FR_{42} (*IF Peakness is High and Height is High and Spacing is High THEN the Confidence value is High*) is the winner.

The Mamdani fuzzy method of the MATLAB tools is also used to compare the inference results under the same conditions (same inputs, same linguistic values, and same ranges). As shown in Fig. 7(a), the fuzzy rules of peakness are aggregated and defuzzied to have a crisp value of Peakness = 0.76. By calculating the centroid, which indicates the rule FR_8 is the winner. In reference to the consequent proposition of FR_8, "Peakness is High" is thus inferred for peakness FPN model. In Fig. 7(b) a crisp value of Height = 0.761, where the rule FR_{23} is the winner. In reference to the consequent proposition of FR_{23}, "Height is High" is thus inferred for Height FPN model. In Fig. 7(c) a crisp value of Spacing = 0.826 is calculated, which indicates the rule FR_1 is the winner. In reference to the consequent proposition of FR_1, "Spacing is High" is

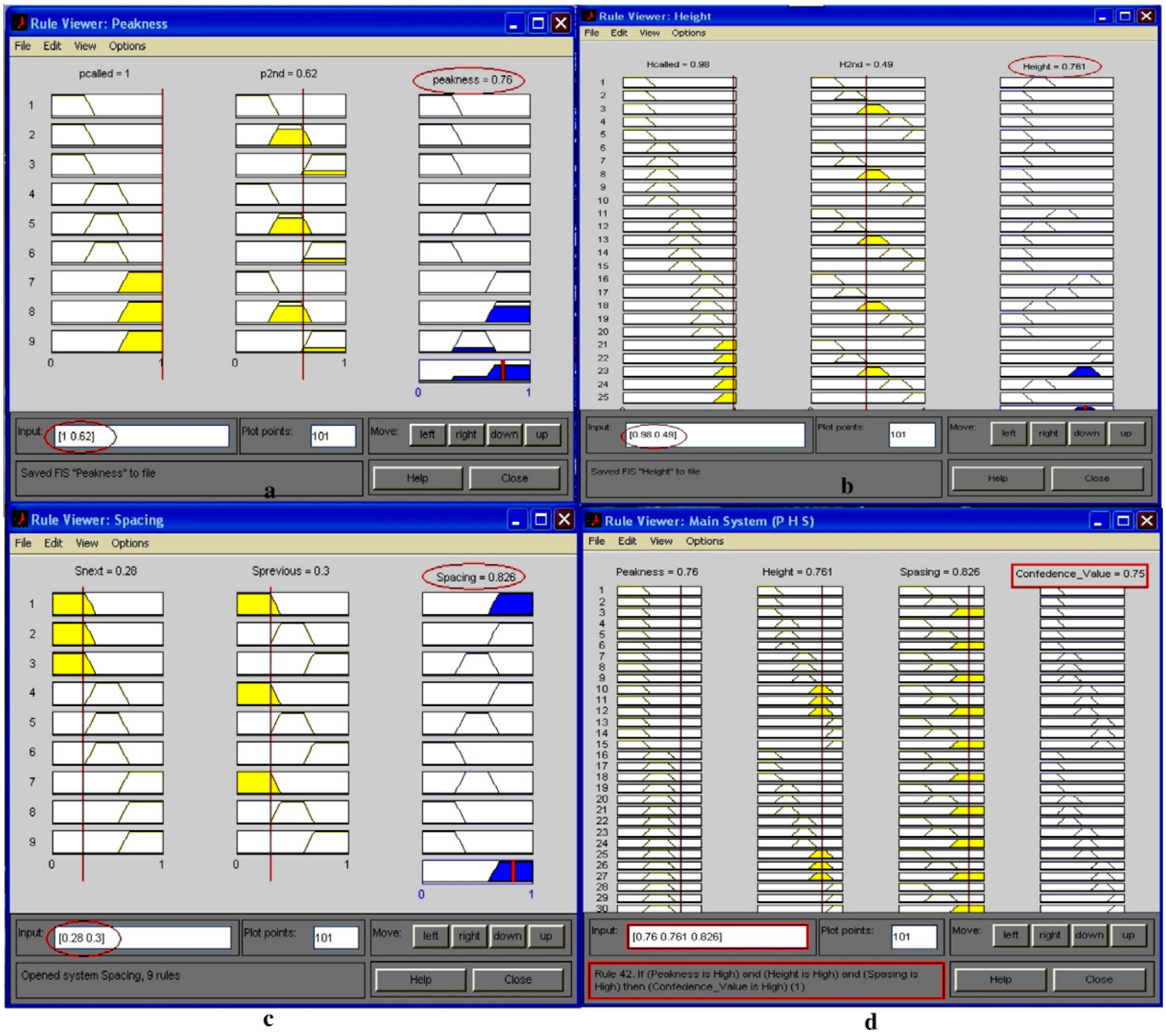

Figure 7 Final decision of (a) a peaknees model, (b) a height model, (c) a spacing model, and (d) a confidence value.

Table 2 Confidence values for the bases called.

	A	T	C	T	C	G
Peakness	0.819	0.826	0.5	0.5	0.569	0.76
Height	0.75	0.925	0.192	0.75	0.35	0.761
Spacing	0.826	0.826	0.826	0.826	0.826	0.826
Confidence	0.75	0.925	0.124	0.75	0.35	0.75

thus inferred for Spacing FPN model. In Fig. 7(d) a crisp value of confidence value = 0.75 is calculated.

Through the comparative study between the methods with the inferred results, both methods have the same reasoning outcomes. While the FPN model clearly shows this distinction in its confidence value, *Phred's* quality value provides exactly the opposite. This shows an inconsistency in the assignment of confidence values or quality values by *Phred*. The Table 2 shows the output data of the all variables and confidence values for the bases called. Using our FPN model, we can say that similar observations can be made for other bases. Comparing the results of the based-calling estimation between the FPN model and the FL model, the similarity that we have discovered is that they both have a same results and a high level of agreement for the based-calling.

6. Conclusion

This paper, introduce a FPN model for fuzzy rule based reasoning. The fuzzy set theory and the fuzzy production rule method are used to establish the fuzzy rules for the confidence value prediction of the bases called in DNA sequencing. This includes the transformation of fuzzy rules into FPN, together with their reasoning. The motivation for using fuzzy Petri nets models is the ability to translate numeric data into linguistic constructs that can then be easily converted into testable hypotheses. It is also worth remarking that the quality values assigned by fuzzy Petri net to determine confidence values for bases called in DNA sequencing are much more informative. We have shown here, that the FPN model is appropriate and can reach the same accuracy performance of available software. The validation was achieved by comparing the results obtained with the FPN model and fuzzy logic using the MATLAB Toolbox; both methods have the same reasoning outcomes. It verifies that the confidence value of the bases called can be successfully reasoned by the proposed FPN model.

Acknowledgement

The authors would like to acknowledge Dr. S. Siddik, Simulation Engineering and my colleagues for their good advices to this paper.

References

Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pacific Symposium on Biocomputing 99, 17–28.

Berno, A., 1996. A graph theoretic approach to the analysis of DNA sequencing data. Genome Research 6 (2), 80–91.

Bostan-Korpeoglu, B., Yazici, A., 2007. A fuzzy Petri net model for intelligent databases. Data & Knowledge Engineering 62, 219–247.

Chen, S.M., Ke, J.S., Chang, J.F., 1990. Knowledge representation using fuzzy Petri nets. IEEE Transactions on Knowledge and Data Engineering 2 (3), 311–319.

Chen, T., He, H.L., Church, G.M., 1999. Modeling gene expression with differential equations. Pacific Symposium on Biocomputing' 99, 29–40.

Ewing, B., Hillier, L., Wendl, M., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8, 175–185.

Fitch, JP., Sokhansanj, B., 2000. Genomic engineering moving beyond DNA sequence to function. Proceedings of the IEEE 88, 1949–1971.

Fujita, S., Matsui, M., Matsuno, H., Miyano, S., 2004. Modeling and simulation of fission yeast cell cycle on hybrid functional Petri net. IEICE Transactions on Fundamentals of Electronics. Communication and Computer Sciences 87 (11), 2919–2928.

Human Genome Project Information, URL: <http://www.ornl.gov/hgmis/>.

Husmeier, D., 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19, 2271–2282.

Liang, S., Fuhrman, S., Somogyi, R., 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing 3, 18–29.

Lukas, W., Ralf, Z., 2008. Intuitive modeling of dynamic systems with Petri nets and fuzzy logic. German Conference on Bioinformatics 2008 (136), 106–115.

Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man–Machine Studies 7 (1), 1–13.

Matsuno, H., Fujita, S., Doi, A., Nagasaki, M., Miyano, S., 2003. Towards Biopathway Modeling and Simulation. In: Proceedings of ICATPN, pp.3–22.

Matsuno, H., Doi, A., Nagasaki, M., Miyano, S., 2000. Hybrid Petri net representation of gene regulatory network. Pacific Symposium on Biocomputing 5, 338–349.

Negnevitsky, M., 2002. Artificial Intelligent–A Guide to Intelligent Systems. Addison-Wesley, New York.

Novak, B., Csikasz-Nagy, A., Gyorffy, B., Chen, K., Tyson, J.J., 1998. Mathematical model of the fission yeast cell cycle with checkpoint controls at the G1/S, G2/M and metaphase/anaphase transitions. Biophysical Chemistry 72, 185–200.

Qu, W., Shirai, K., 2003. Belief learning in certainty factor model and its application to text categorization. In: Proceedings of the 2003 Joint Conference of the Fourth Inter. Con. on Infor., Comm. and Signal Processing and Fourth Pacific-Rim Conference on Multimedia, vol. 12, pp. 1192–1196.

Raed, I.H., Ahson, S.I., 2010. Designing Genetic Regulatory Networks Using Fuzzy Petri Nets Approach. (IJAC). Springer-Verlag, Berlin Heidelberg, issue: 7, vol. 3, pp. 403–412.

Raed, I.H., Ahson, S.I., 2010. A new approach for modeling gene regulatory networks using fuzzy Petri nets. Journal of Integrative Bioinformatics 7 (113), 1–16.

Ressom, H., Natarjan, P., Varghese, R.S., Musavi, M.T., 2005. Applications of fuzzy logic in genomics. Journal of Fuzzy Sets and Systems 152, 125–138.

Vohradsky, J., 2002. Neural networks model of gene expression. The FASEB Journal 15, 846–854.

Zadeh, L.A., 1975. The concept of linguistic variable and its applications to approximate reasoning-II. Information Sciences 8, 301–357.

Zadeh, L.A., 2002. Precisiated natural language – toward a radical enlargement of the role of natural languages in information processing, decision and control. In: Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP'02), vol. 1, pp. 1–3.