A Novel Application of Logic Petri Nets to Rule-Based Reasoning of DNA Base-Calling

Jing SHA*1, Zhennan ZHANG2

1College of Computer Science, Shandong University of Science and Technology, Qingdao, Shandong 266590, China.
2Computer Teaching Section, Weifang Medical University, Weifang, Shandong 261053, China.

Abstract — The importance of DNA sequencing necessitates the efficient automation of identification of base sequences from traces generated by existing sequencing machines, a process referred to as DNA base-calling. The proposed approach brings DNA bases called within the framework of a new tool called Logic Petri Nets, LPNs. This study adopts a pattern recognition technique to minimize inaccuracies in DNA base-calling. The input features in the LPN model of the base called, can be formulated as uncertain tokens to determine the confidence value. Reasoning rules integrated in a logical input expression determine the firing of logical input transitions. Lastly, a new algorithm is presented to mechanize the computation the approach.

Keywords - base-calling; logic Petri net; sequencing; rule-based reasoning

I. INTRODUCTION

The ability to decipher the genetic code of different species leads to significant future scientific achievements in important areas, including medicine and agriculture. Different analysis tools are being developed to detect and understand the phenomena of gene regulation and physiological functions, and to assess the quality of a genomic sequence. DNA sequencing does not only have significant scientific implications in the context of deciphering the fundamental code of life, but also represents an enormous opportunity to improve the well-being of humankind by ushering in a new era of personal or precision medicine. A DNA molecule has a double helical structure with two intertwined chains that comprise complementary nucleotide strands, where A bonds with T and C pairs up with G [1]. Information derived from the genomic sequence is likely to contribute enormously to medical advances, such as more accurate diagnosis of genetic diseases, improved drug design to target specific genes causing certain diseases, and gene therapy through replacement of defective genes. A major challenge of modeling biological systems is that conventional methods based on physical and chemical principles require data that are 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 ratio rather than physical quantity) [2]. Various kinds of models have been studied to express biological systems, such as differential equations [3], Boolean networks [4], Petri nets (PNs) [5], [6], [7], and artificial neural networks [8]. The abovementioned papers are dedicated to the applications of different methods to genetic networks and show that these methods are suitable to model special molecular biological systems.

Several systems have been designed in the past two decades to facilitate and automate DNA base-calling. The growth in the number of genomic datasets generated with tools, such as high-throughput DNA sequencing machines and DNA microarrays, has created a critical requirement for resources that facilitate the interpretation of large-scale biological data. New mathematics and novel methodologies are required to contribute to the conceptual or complex theoretical framework in which biologists study organisms. This study presents a novel tool referred to as logic Petri net (LPN), which can satisfy the requirement for a conceptual framework and provide a systematic and unbiased way to perform this transformation.

II. LOGIC PETRI NETS

A. Formal Definition of Logic Petri Nets

A LPN [9], [10] is a high-level PN. LPNs can well describe and analyze batch processing functions and passing value indeterminacy in cooperative systems. Two kinds of logic transitions occur: logic input and logic output transitions. Logic transitions are restricted by logic expressions that can efficiently describe batch processing function and passing value indeterminacy, such as the indeterminacy of arrival calls or actual trading quantities in electronic commerce.

Given that only logic expressions are studied to predict the confidence values of genes called in DNA sequencing, the capacity of each place is limited to 1. The definitions of PNs and LPNs are presented below.

Definition 1: \( PN = (P, T, F, M_0) \) is a PN if and only if:

1. \( P \) is a finite non-empty set of places;
2. \( T \) is a finite non-empty set of transitions, and \( P \cap T = \Phi \);
3. \( F \subseteq (T \times P) \cup (P \times T) \) is a set of arcs, i.e., flow relations; and
4. \( M_0 : P \rightarrow \{0,1\} \) is the initial marking.

In the graphic representation, places are drawn as circles or ellipses, while transitions are illustrated as bars. The flow relations between the nodes are represented as directed arcs, while the tokens of the markings are shown as dots inside the places. The following symbols are used to represent the pre-
and post-set of a node. \( x \in P \cup T^*: x = \{ y \mid (x, y) \in E \}, \)
\( x^* = \{ y \mid (x, y) \in F \}. \) \( R(M_o) \) is a set of all markings reachable from \( M_o \). A transition \( t \) is enabled if each \( pe^t \) contains one token at the present marking \( M \in R(M_o). \) If \( t \) is enabled, it can fire and a new marking \( M' \) is generated from the current marking \( M \), as represented by \( T \). \( \alpha \) is the certainty factor (CF) of a transition. \( \beta \) is enabled and \( \gamma \) is the function that assigns a token value to each place; and \( \delta \) has uncertain state of input and output places. Figure 1 shows an LPN model. In the model, \( t_1 \) is a logic input transition. The firing of \( t_1 \) is restricted by the logic expression \( f_1(t_1) = p_1 \lor p_2; \) \( t_3 \) is a logic output transition, the firing is restricted by the logic expression \( f_3(t_3) = p_6 \lor (p_7 \lor p_8) \). From the logic values of the input places, the value of the logic expression is 0, 0. 0. \( \beta\) is a logic \( PN \) if and only if

(1) \( P \) is a finite set of places; 
(2) \( T = T_D \cup T_I \cup T_O \) is a finite set of transitions, \( T \cup P \neq \Phi \), \( \forall t \in T \cup T_D, t \cap T_I = \Phi \), and \( P, T_D, T_I, T_O \) are disjoint with one another, where 
(a) \( T_D \) denotes a set of traditional transitions; 
(b) \( T_I \) denotes a set of logic input transitions, and all input places of \( t \) are restricted by a logic input expression \( f_I(t) \); and 
(c) \( T_O \) \( \forall t \in T_O \) denotes a set of logic output transitions, and \( \forall t \in T_O \); all output places of \( t \) are restricted by a logic output expression \( f_O(t) \). 

(3) \( F \subseteq (P \times T)(T \times P) \) denotes a finite set of arcs; 
(4) \( I \) is a logic input function, and \( \forall t \in T_I, I(t) = f_I(t) \) is a logic input expression; 
(5) \( O \) is a logic output function, and \( \forall t \in T_O, O(t) = f_O(t) \) is a logic output expression; 
(6) \( M : P \rightarrow \{0, 1\} \), is a marking function, \( \forall p \in P, M(p) \) is the number of tokens in \( p \); 
(7) The rules of the trigger for the transition are

(a) \( \forall t \in T_D \), the rule is similar to the transition in traditional PN; 
(b) \( \forall t \in T \), \( I(t) = f_I(t) \), if \( f_I(t) \mid_{\Delta t} = T \), that is, \( \ast \) satisfies a logic input expression \( f_I(t) \) at \( M \), then \( t \) is enabled at \( M \); if \( t \) is enabled, it can be fired and generates a new state \( M' \), \( \forall p \in \ast t: M(p) = 0；\) \( \forall p \in \ast t \setminus \ast t: M(p) = M(p) \); 
(c) \( \forall t \in T_O: O(t) = f_O(t), \) if \( \forall p \in \ast t: M(p) = 1 \), then \( M' \); \( \forall M \geq T; \) then 
\( \forall p \in \ast t: M(p) = M(p) - 1；\) 
\( \forall p \in \ast t \setminus \ast t: M(p) = M(p) \); 
\( \forall p \in \ast t: f_O(t) \mid_{\Delta t} = T \), thus, \( \ast t \) must satisfy a logic expression \( f_O(t) \) at \( M \). 

(8) \( \alpha : P \rightarrow [0, 1] \) is the function that assigns a token value between 0 and 1 to each place; and 
(9) \( \beta : T \rightarrow [0, 1] \) is the certainty factor (CF) of a transition. 

In LPNs, logic input and output transitions are restricted by the logic input expression \( f_I(t) \) and output expression \( f_O(t) \). The rules to trigger logic transitions present the
III. METHODS AND MODELING OF DNA BASES CALLED

In many vital phenomena, genomes represent the most basic unit of information. Genomes convey genetic information from ancestors to descendants, and accumulated polymorphisms are significant in the evolution of organisms. Human genome refers to the heredity information encoded in the DNA of Homo sapiens and stored in 23 pairs of chromosomes located in the cell nucleus. A DNA strand consists of four nucleotide bases: adenine (A), cytosine (C), thymine (T), and guanine (G) [11].

Over a decade after the completion of the Human Genome Project in 2003, effort is now devoted to the development of next-generation DNA sequencing technologies that can meet the “$1,000 genome” goal set by the National Institute of Health.

The Mamdani fuzzy inference system [12] was proposed as the first attempt to control a steam engine and boiler combination through 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 the LPN that can describe logic inputs that control the firing of transitions. According to the raw data of the trace, a DNA sequencing analysis software, such as base-callers, estimate the most likely set of bases in the original DNA sample. The PHRED program [13] is currently the most commonly used base-caller. Reference [14] provided a logic algorithm to predict the confidence values for each base called in DNA sequencing. This technique uses three kinds of information to assess the accuracy of DNA bases called, namely, height, peak, and spacing of the base considered and the next candidate. In the present study, only the original reasoning rules are considered to predict the confidence value of the bases called.

A. LPN model

As illustrated in Fig. 4, the model is used to describe the inference reasoning system. The figure presents the contents of the reasoning rules in the model. The properties of the proposition set of places, and the logic input and output transitions are described as follows:

1) I1, I2 represent the “input data.” Each place has a token value to provide the normalized value of the base considered.

2) The first transitions, D1, D2, represent the distribution function transition of the input information. These two transitions are logic output transitions. The logic output functions are \( f_{D1}(t) = p_1 \lor p_2 \lor \ldots \lor p_j \) and \( f_{D2}(t) = p_1 \lor p_2 \lor \ldots \lor p_j \).

3) Places L1, M1, H1, L2, M2, and H2 are places that hold the processed data to be compared.

4) Transitions t1, t2, \ldots, t9 are the logic input transitions, and the logic input functions of the transitions are \( f_{t1}(t) = L_1 \land L_2 \), \( f_{t2}(t) = L_1 \land M_2 \), \ldots, \( f_{t9}(t) = H_1 \land H_2 \). These transitions provide the “minimum” operation for the values of the two antecedent places.

5) Places p1, p2, \ldots, p9: represent places that hold the values of the “minimum” operation.

6) Transitions t10, t11, t12 are the logic input transitions, and the logic input functions of the transitions are \( f_{t10}(t) = p_1 \lor p_2 \lor p_3 \), \( f_{t11}(t) = p_1 \lor p_2 \lor p_3 \), \( f_{t12}(t) = p_4 \lor p_5 \lor p_6 \). These transitions provide the “maximum” operation for the value of the antecedent places.
This study presents an LPN model for rule-based reasoning. Fuzzy set theory and fuzzy production rule method are used to establish the rules for the confidence value prediction of the bases called in DNA sequencing. This process includes the transformation of rules into FPN, together with the reasoning for the rules. Notably, the quality values assigned via LPN to determine the confidence values for bases called in DNA sequencing are much more informative. The algorithm presented in the last section is further simulated using tools to realize the study.

ACKNOWLEDGMENT

This work is sponsored by Science and technology development plan of Shandong Province (No.2012G0020120), Shandong Province Higher Educational Science and Technology Program under Grant (No.J13LN18, No.J15LN19).

REFERENCES