Modeling, Learning and Simulating Biological Cells with Entity Grammar

Yun Wang^{1,*} Rao Zheng², and Yan-Jiang Qiao¹

¹ Beijing University of Chinese Medicine, Beijing, 100102, China

² Beijing University of Chemical Technology, Beijing, 100029, China
yonwangpku@yahoo.com.cn

Abstract. In recent years, whole cell modeling approaches, which combine existing knowledge and machine learning results, have received considerable attention. These approaches are potentially very efficient for simulating and analyzing physiological function of cells. In this work, entity grammar system is proposed as formalism for knowledge representation and multi-strategy learning techniques in systems biology. Modeling biological cells with entity grammar starts from the simple grammatical models. Integrating the simple models into a cooperating entity grammar of cell facilitates real-time model learning and updating. This makes a difference with many other formalisms to build preset models. The scheme of such a platform is described in the paper and the possible applications are discussed. The proposed formalism method is open to all reasoning paradigm and can be used for studying biological complex systems.

Keywords: entity grammar system, systems biology, complex system.

1 Introduction

Computational models in biology connect molecular mechanisms to the physiological properties of cell, which is important to integrate the knowledge of cells. There have been many previous efforts to design simulators for general purpose such as GEPASI[1], E-CELL[2], Virtual Cell[3], BioDrive[4], Cellerator[5] etc. The involved formalisms include differential equations, Boolean networks, logical/graph-based (LG), Bayesian Networks, Petri-nets, π -calculus, etc. Most of the formalisms are based on the settled knowledge about cells, which deal with the problems of modeling from knowledge. To solve the problems of knowledge discovery from data, machine learning techniques are required. However, in most conditions, modeling from knowledge and knowledge discovery from data are concurrent. Combination of modeling and simulation procedures with machine learning techniques will promote and simplify both the process of model validation and data integration. Although much endeavor has been done in this direction, modeling, learning and simulating complex systems require more flexible formalism, which is also one of the main tasks in systems biology.

* Corresponding author.

To meet the need for new modeling formalism, learning strategies, simulation platforms and their integration, we present the formalism of Entity Grammar System (EGS) as a potential choice for these tasks.

2 Entity Grammatical Model of Biological Cell

In this section, we will recall the basic definitions of EGS. For more details, please refer to reference [6].

An entity grammar G is a quintuple, $G = (V_N, V_T, F, P, S)$, where, V_N is finite set of non-terminal symbols, V_T is finite set of terminal symbols, and $V_N \cap V_T = \Phi$, F is finite set of operations, $F = \{f_i \mid f_i : (E(V, F))^n \rightarrow E(V, F), 1 \leq i \leq m, m, n \in \mathbb{N}\}$, where $V = V_N \cup V_T$, P is a finite set of productions $\alpha \rightarrow \beta$ with $\alpha \in E^+(V, F)$ and $\beta \in E(V, F)$, S is the start entities.

A cooperating entity grammar of n EGSs is denoted by

$$G = (g_{V_N}(V_{N,1}, \dots, V_{N,n}), g_{V_T}(V_{T,1}, \dots, V_{T,n}), g_F(F_1, \dots, F_n), g_P(P_1, \dots, P_n), g_S(S_1, \dots, S_n)) \quad (1)$$

Where g is the cooperating functions of alphabets, sets of operations, sets of production rules and sets of start entities.

To model biological system, the cells can be represented by several cooperating systems, including the DNAs, RNAs, proteins, membrane and small molecules. The subsystems cooperate through their interactions in different levels.

Suppose we have the following entity grammatical models:

$$G_{\text{DNA}} = (V_{\text{DNA}}, F_{\text{DNA}}, P_{\text{DNA}}, S_{\text{DNA}}) \quad (2)$$

$$G_{\text{RNA}} = (V_{\text{RNA}}, F_{\text{RNA}}, P_{\text{RNA}}, S_{\text{RNA}}) \quad (3)$$

$$G_{\text{protein}} = (V_{\text{protein}}, F_{\text{protein}}, P_{\text{protein}}, S_{\text{protein}}) \quad (4)$$

According to the basic definition of cooperating entity grammar, the cooperating entity grammatical model of DNAs, RNAs and proteins can be denoted as

$$G = \left(g_V(V_{\text{DNA}}, V_{\text{RNA}}, V_{\text{protein}}), g_F(F_{\text{DNA}}, F_{\text{RNA}}, F_{\text{protein}}), g_P(P_{\text{DNA}}, P_{\text{RNA}}, P_{\text{protein}}), g_S(S_{\text{DNA}}, S_{\text{RNA}}, S_{\text{protein}}) \right) \quad (5)$$

In general, the alphabets of the three grammars are different, the function g_V could be concreted as the union operation, denoted by

$$g_V(V_{\text{DNA}}, V_{\text{RNA}}, V_{\text{protein}}) = V_{\text{DNA}} \cup V_{\text{RNA}} \cup V_{\text{protein}} \quad (6)$$

The function g_F is the operation on the organizers of the three grammars. Overlap among the elements of the three sets of operations requires deletion of the iterative operations and addition of new operations describing new structures. If the set of new operations is denoted by $F_{\text{DNA-RNA-protein}}$, the function g_F can be denoted by

$$g_F(F_{DNA}, F_{RNA}, F_{protein}) = F_{DNA} \cup F_{RNA} \cup F_{protein} \cup F_{DNA-RNA-protein} \quad (7)$$

Similarly, function g_P can be denoted by

$$g_P(P_{DNA}, P_{RNA}, P_{protein}) = P_{DNA} \cup P_{RNA} \cup P_{protein} \cup P_{DNA-RNA-protein} \quad (8)$$

Where, $P_{DNA-RNA-protein}$ is the rewriting procedures involving at least two kinds of molecules. The choice of the start entities of the cooperating entity grammar depends on the purpose of research. The elements of the start entities could be DNAs, RNAs, proteins or their complexes, only if they can be described by the operations in g_F . Such a modeling method in EGS provides the possibility for establishing the models of meta-systems.

3 Entity Grammatical Model-Based Learning of Cells

In this section, we present the basic ideas of knowledge representation and learning strategy implementation based on the formalism of EGS. The approach represents the statements of knowledge as entity grammars. The collection of knowledge is the cooperating entity grammars. Entity grammatical model-based learning includes four transforms: entity generalization (induction), entity specialization (deduction), entity similization (analogy) and entity dissimilization (analogy). The process of how knowledge transmutations modify a statement in EGS can be realized in the environment of multi-programming paradigms (*e.g. Mathematica*).

The kernel of the platform includes a knowledge library in the form of entity grammars and a process of knowledge collection to build the knowledge library(Fig. 1). It provides the mechanisms of knowledge integration and the interface for knowledge input. The simulator of the platform is both the interface for query input and the channel for communicating with the kernel.

1. Update the knowledge library with new information; 2. Search and retrieve knowledge; 3. Replace; 4. Induction; 5. Input start entities; 6. Deduction; 7. Analogy; 8. Answers

Fig. 1. The architecture of entity grammar based platform

The original knowledge library is empty. With the input of knowledge from instructor and the observed facts, the model would become more and more powerful. The relationships between input knowledge and initial knowledge mainly include five types: (1) The input being included by initial knowledge would be discarded. (2) The input including initial knowledge library would be updated. (3) The input of new information would be added to the initial knowledge library. (4) For the input similar to initial knowledge, the kernel will use the induction approaches to gain the new knowledge from the input and initial knowledge and then update the library. (5) The input contradicting with initial knowledge will be reevaluated and accepted if it is more credible than initial ones. Otherwise, the input will be discarded.

The knowledge library in this system is an entity grammar with the start entities being an empty set. The definition of the start entities is left to users according to their interested problems. The simulator will input them to the kernel for simulation.

4 Concluding Remarks

In this paper, entity grammar system is presented as formalism approach for integrating diverse data and knowledge in biological cells. The platform for modeling, learning and simulating biological cells using entity grammars is described. This formalism is open to all knowledge from different cells and most of the reasoning paradigms, including induction, deduction and analogy. The cooperating entity grammar provides the methods for real-time model learning and updating. This make a difference with many other formalisms to build preset models. The approach can be used to solve problems in biological complex systems. With the development of the related concrete techniques, the formalism approach will facilitate whole cell modeling and reasoning.

Acknowledgements. This work was financially supported by the Natural Science Foundation of China (No. 30500643).

References

1. Mendes, P.: GEPASI: A Software Package for Modeling the Dynamics, Steady States and Control of Biochemical and Other Systems. *Comput. Appl. Biosci.* (1993) 563-571
2. Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Clyde, A., Hutchison, C.A.: E-CELL: Software Environment for Whole-cell Simulation. *Bioinformatics.* (1999) 72-84
3. Schaff, J., Loew, L.M.: The Virtual Cell. *Proceedings of Pacific Symposium on Biocomputing '99.* (1999) 228-239.
4. Kyoda, K.M., Muraki, M., Kitano, H.: Construction of a Generalized Simulator for Multi-Cell Organisms and its Application to SMAD Signal Transduction. *Pacific Symp. Biocomputing.* (2000) 314-325
5. Shapiro, B.E., Levchenko, A., Mjolsness, E.: Automatic Model Generation for Signal Transduction with Applications to MAP-Kinase Pathways. in *Foundations of Systems Biology*, Kitano, H. (ed.), MIT Press, Cambridge, MA. (2001)
6. Wang, Y.: Entity Grammar Systems: A Grammatical Tool for Studying the Hierarchal Structures of Biological systems. *Bull. Math. Biol.* (2004) 447-471