A Semantic Expert System for the Evolutionary Design of Synthetic Gene Networks

Vitoantonio Bevilacqua
DEI – Polytechnic of Bari
Via Orabona, 4 - 70125
Bari - Italy
00390805963326
vitoantonio.bevilacqua@poliba.it

Paolo Pannarale
DEI – Polytechnic of Bari
Via Orabona, 4 - 70125
Bari - Italy
00390805963326
p.pannarale@gmail.com

ABSTRACT

This work tries to cope with the standardization issue by the adoption of model exchange standards like CellML, BioBrick standard biological parts and standard signal carriers for modeling purpose. The BioBricks are easily assemblable [1] standard DNA sequences coding for well-defined structures and functions and represent an effort to introduce the engineering principles of abstraction and standardization in synthetic biology. Web applications as GenoCAD [2] are available and implements an algorithm of syntax check of the circuits designed [3], while some other tools for automatic design and optimization of genetic circuits have appeared [4] and are also specific for BioBrick systems [5]. Our generated models are made of Standard Virtual Parts modular components. Model complexity includes more interaction dynamics than previous works. The inherent software complexity has been handled by a rational use of ontologies and rule engine. The database of parts and interactions is automatically created from publicly available whole system models. We implemented a genetic algorithm searching the space of possible genetic circuits for an optimal circuit meeting user defined input-output dynamics. The tools performing structural optimization usually use stochastic strategies, while those optimizing the parameters or matching the components for a given structure can take advantage of both stochastic and deterministic strategies. In most cases it is however necessary human intervention, for example to set the value of certain kinetic parameters. To our best knowledge no tool exists which does not show a couple of these limitations, then our tool is the only capable of using a library of parts, dynamically generated from other system models available from public databases [6]. The tool automatically infers the chemical and genetic interactions occurring between entities of the repository models and applies them in the target model if opportune. The repository models have to be modeled by a specific CellML standard, the Standard Virtual Parts (SVP) [7] formalism and the components have to be annotated with OWL for unique identifiers. The output is a sequence of readily composable biological components, deposited in *the registry of parts,* and a complete CellML kinetic model of the system.

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2881-4/14/07.
http://dx.doi.org/10.1145/2598394.2598504

Accordingly, a model can be generated and simulated from a sequence of BioBrick, without any human intervention. Actual tools present a moderated degree of accuracy in the prediction of the behavior, principally due to the lack of consideration of many cellular factors. Despite the advances in molecular construction, modeling and fine-tuning the behavior of synthetic circuits remains extremely challenging [8]. We tried to cope with this issue of scalability by means of ontologies coupled with a rule engine [9]. Model complexity includes more interaction dynamics than previous works, including gene regulation, interaction between small molecules and proteins but also protein-protein and post-transcriptional regulation. The domain was described by using Ontology Web Language (OWL) ontologies in conjunction with CellML [10], while complex logic was added by Jess rules [11]. The system has been successfully tested on a single test case and looks towards the creation of a web platform [12].

Categories and Subject Descriptors

D.1.0 [**Programming Techniques**]: general; I.2.m.c [**Artificial Intelligence**]: Evolutionary computing and genetic algorithms; I.2.1 [**Artificial Intelligence**]: Applications and Expert Knowledge-Intensive Systems.

Keywords

Biological system modeling, Design automation, DNA, Expert systems, Genetic algorithms.

1. PROBLEM FORMULATION

The problem of finding the optimal genetic circuit was formulated as the genetic algorithm search of an optimal path in the graph representing the admissible parts compositions. Where the total number of paths is given by:

$$N_P = \frac{1 - N_b^{r-1}}{1 - N_b}. \quad (1)$$

and the required output for the ith variable is:

$$\hat{f}_i(t), \ i = 1...n \ . \quad (2)$$

V is the set of size N_b of the BioBricks composing the palette of components that the solution can be made of and $G=(V,E)$ is a connected directed graph representing the admissible composition of BioBricks. The stimulation protocol for the jth input is:

$$u_j(t), \ j = 1...m \ . \quad (3)$$

t_f defines the simulation horizon, with t belonging to $[0, t_f]$

A path $P = (v_o, \ldots v_r) | (v_{k-1}, v_k)$ belonging to E, for each k

belonging to [1,r] , the fitness function is:

$$\min_{P} \Phi\left(\hat{f}, f^{P}\right) \tag{4}$$

Φ is a measure of how much the output corresponding to a BioBrick system P, approximates the desired output, for all the considered *i* variables from the time 0 to t_f;
and constraints are the occurrences of a particular node vnd, indicating the begin of a new synthetic device must be less than ND_{max} and the total number of nodes must be $R_{min} < r < R_{max}$.

The search algorithm used is a genetic algorithm defined in such a way that the chromosomes represent real biological plasmids. The candidate solutions are transformed into a CellML model by an expert system, called MCE (Model Construction Engine). The engine adopts a hybrid strategy based on OWL and Jess rules. This choice allows the use of the many ontologies created for the world of molecular biology, to exploit the effectiveness of ontologies in the representation of the entities of a domain, but also exploits the flexibility and simplicity of implementation of Jess rules. Once logically reconstructed, the model has to be translated into a CellML model, in terms of components, variables, connections, sums of flows, flow equations, initial values and multiple encapsulations. The CellML simulation service is able to take as input the CellML model and to produce a simulation using a large number of algorithms and the related configuration parameters. Once obtained the similation results, the fitness value can be computed by

$$\Phi = \prod_{i=0}^{n} d(\hat{f}_i, f_i^{P}) \tag{5}$$

productory of the distance for every ith variable, of the time course obtained from the model corresponding to the path P and the target behavior. The distance measured has been defined as follows:

$$d = 10^{-\frac{\sum_{t=0}^{t_f} \left\| \log\left(\hat{f}_i(t)\right) - \log\left(\max\left(\alpha \cdot \max_t\left(\hat{f}_i\right), f_i^{P}(t) \right) \right) \right\|}{t_f}}. \tag{6}$$

The fitness function belongs to the interval [0,1] and penalizes those circuits that do not meet all the objectives simultaneously.

2. FINAL RESULTS AND CONCLUSIONS

The repressilator is an oscillatory synthetic gene network designed by Elowitz et al. [12], its model of the CellML repository was not represented accordingly to the SVP abstraction and an appropriate conversion has been necessary. The existing SVP templates were not sufficient for its modeling and a new repressible promoter template has been introduced. Its flux rate is given by

$$JRNA = \frac{\left(k0 + \frac{k \cdot Km^{n}}{Km^{n} + tf^{n}} \right) \cdot E^{9}}{vol \cdot E^{-15} \cdot a}, \tag{7}$$

where vol is the volume expressed in femtoliters, a is the Avogadro's constant, k is the promoter efficiency expressed as PoPs, tf is the repressor nanomolar concentration; the flow is expressed as nanomolars per second. The SVP ontology has been modified to add the new repressible promoter entity, as well as we have added a rule that identifies these entities as those that implement the newly defined template. Finally, the rule that connects the variable expressing the concentration of the repressor with the corresponding value in the model of the promoter has been created. The output of the model obtained by the Construction Engine Model includes 30 components, 23 import, 6 equations, 24 variables, 56 connections and 2 units and takes some hours to be developed by hand but a few minutes with the aid of ARChITeCt. Architect's main objective is to provide a tool for the automated design and modeling of genetic circuits and at the same time promote the kinetic characterization of reusable biological components. To do this a framework that facilitates sharing, based on established standards such as CellML or SBML, and provide a real benefit to the user, is needed. The OWL implications are themselves expressed as Jess rules. In this way it is possible to combine in a single reasoner the advantages of OWL and the possibility of using rules for particular tasks, without expressivity limitations, obviously at the cost of no computational guarantees.

3. REFERENCES

[1] V. Bevilacqua, et al., "Comparison of data-merging methods with svm attribute selection and classification in breast cancer gene expression," BMC Bioinformatics, 2012, vol. 13, no. Suppl 7, p. S9.

[2] Y. Cai, et al., "GenoCAD for iGEM: a grammatical approach to the design of standard-compliant constructs." Nucleic acids research, May 2010, vol. 38, no. 8, pp. 2637–44.

[3] J. Kearney, et al., "Software complexity measurement," Communications of the ACM, 1986, vol. 29, no. 11, pp. 1044–1050.

[4] J. Kelly et al., "Tools and Reference Standards Supporting the Engineering and Evolution of Synthetic Biological Systems". Massachusetts Institute of Technology, Biological Engineering Division, 2008.

[5] J. R. Kelly, et al., "Measuring the activity of BioBrick promoters using an in vivo reference standard." Journal of biological engineering, Jan. 2009, vol. 3, p. 4.

[6] M. Kaern, et al. , "The engineering of gene regulatory networks." Annual review of biomedical engineering, Jan. 2003, vol. 5, pp. 179–206.

[7] G. Karlebach et al., "Modelling and analysis of gene regulatory networks." Nature reviews. Molecular cell biology, , Oct. 2008, vol. 9, no. 10, pp. 770–80.

[8] T. Knight, "Idempotent vector design for standard assembly of biobricks," DTIC Document, Tech. Rep, 2003.

[9] T. Knight Jr, et al. "Biojade: a design and simulation tool for synthetic biological systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[10] F. Licciulli et al., "Mblabdb: a social database for molecularbiodiversity data," EMBnet.journal, 2012, vol. 18, no. B.

[11] T. Shimayoshi, et al. "A Method to Support Cell Physiological Modelling Using Description Language and Ontology," IPSJ Digital Courier, 2006, vol. 2, no. 1, pp. 726–735.

[12] M. Aylward, et al. "Bugbuster: Computational design of a bacterial biosensor," in iGEM, 2008.