

Grammatical Evolution for the Discovery of Petri Net Models of Complex Genetic Systems

Jason H. Moore and Lance W. Hahn

Program in Human Genetics, Vanderbilt University, Nashville, TN, USA 37232-0700
{Moore,Hahn}@phg.mc.Vanderbilt.edu

Abstract. We propose here a grammatical evolution approach for the automatic discovery of Petri net models of biochemical systems that are consistent with population level genetic models of disease susceptibility. We demonstrate the grammatical evolution approach routinely identifies interesting and useful Petri net models in a human-competitive manner. This study opens the door for hierarchical systems modeling of the relationship between genes, biochemistry, and measures of health.

Petri nets are a type of directed graph that can be used to model discrete dynamic systems. The goal of this study was to develop a grammatical evolution (GE) strategy for the automatic discovery of Petri net (PN) models of biochemical systems that are consistent with gene-gene interactions that increase susceptibility to human disease. Understanding the relationship between genes, biochemistry, and measures of health is an important endeavor in the domain of human genetics. We first summarize the PN modeling strategy and then briefly present the grammar used and the results.

The goal of identifying PN models of biochemical systems that are consistent with observed population-level gene-gene interactions is accomplished by developing PN that are dependent on specific genotypes from two (or more) genetic variations. Here, we make firing rates of transitions and/or arc weights genotype-dependent yielding different PN behavior. Each PN model is related to the genetic model using a threshold model. With the threshold model, it is the concentration of a substance that is related to the risk of disease. For each model, the number of tokens at a particular place is recorded and if they exceed a certain threshold, the appropriate risk assignment is made. If the number of tokens does not exceed the threshold, the alternative risk assignment is made. The high-risk and low-risk assignments made by the discrete threshold from the output of the PN can then be compared to the high-risk and low-risk genotypes from the genetic model. A perfect match indicates the PN model is consistent with the gene-gene interactions observed in the genetic model. Here, the fitness function of the GE algorithm is proportional to the number of high-risk and low-risk assignments incorrectly made.

We developed a grammar for PN in Backus-Naur Form (BNF). Nonterminals form the left-hand side of production rules while both terminals and nonterminals can form the right-hand side. A terminal is essentially a model element while a nonterminal is the name of a production rule. For the PN models, the terminal set includes, for

example, the basic building blocks of a PN places, arcs, and transitions. The nonterminal set includes the names of production rules that construct the PN. For example, a nonterminal might name a production rule for determining whether an arc has weights that are fixed or genotype-dependent. We show below in (1) the production rule that is executed to begin the model building process.

(1) <root> ::= <pick_a_gene> <pick_a_gene> <net_iterations> <expr> <transition> <place_noarc>

When the initial <root> production rule is executed, a single PN place with no entering or exiting arc (i.e. <place_noarc>) is selected and a transition leading into or out of that place is selected. The arc connecting the transition and place can be dependent on the genotypes of the genes selected by <pick_a_gene>. The nonterminal <expr> is a function that allows the PN to grow. The production rule for <expr> is shown below in (2). Here, the nonterminal (0, 1, 2, or 3) selected in the right-hand side of the production rule is determined by a combination of bits in the genetic algorithm chromosome.

(2)	<expr> ::= <expr> <expr>	0
	<arc>	1
	<transition>	2
	<place>	3

The base or minimum PN that is constructed using the <root> production rule consists of a single place, a single transition, and an arc that connects them. Multiple calls to the production rule <expr> by the genetic algorithm chromosome can build any connected PN. In addition, the number of times the PN is to be iterated is selected with the nonterminal <net_iterations>. Many other production rules control the arc weights, the genotype-dependent arcs and transitions, the number of initial tokens in a place, the place capacity, etc. All decisions made in the building of the PN model are made by each subsequent bit or combination of bits in the genetic algorithm chromosome.

The GE algorithm was run a total of 100 times for each of two hypothetical nonlinear gene-gene interaction models. For each genetic model, GE always yielded a PN model that was consistent with the high-risk and low-risk assignments for each combination of genotypes with no classification error. Most PN models consisted of one place, two arcs, and two transitions. We found that there was a clear preference for making arcs, rather than transitions, genotype-dependent. Understanding how interactions at the biochemical level manifest themselves as interactions among genes at the population level, will provide a basis for understanding the role of genes in diseases susceptibility.

This work was supported by National Institutes of Health grants HL65234, HL65962, GM31304, AG19085, and AG20135.