
Turning Genes Off and On: Using Genetic Algorithms with Complexity-Based Fitness for Model Selection in Ecology

James P. Hoffmann

Department of Botany &
Agricultural Biochemistry
University of Vermont
Burlington, VT 05405

Chris D. Ellingwood

Department of Botany &
Agricultural Biochemistry
University of Vermont
Burlington, VT 05405

Osei M. Bonsu

Department of Mathematics
& Statistics
University of Vermont
Burlington, VT 05405

Daniel E. Benti

Department of Mathematics
& Statistics
University of Vermont
Burlington, VT 05405

Abstract

This paper describes experiments with a genetic algorithm that combines parsimony with a novel gene regulation mechanism to carry out model selection. In effect, the GA orchestrates a competition among a community of models. Parsimony is implemented via the Akaike Information Criterion, and gene regulation uses a modulo function to overload the gene values. The approach is shown to be successful with polynomial models and complex biological simulation models, even when Gaussian noise is added to the data.

1 INTRODUCTION

Model selection is central to our modern method of doing science. Guiding principles for selecting the best model among a set of competing models have become increasingly sophisticated since William of Occam postulated his now famous qualitative criterion: the simplest model that adequately describes the empirical data is usually the correct one (Occam's razor). Occam's emphasis on simplicity (parsimony) is implemented in the quantitative model-selection methods we use today, including classical hypotheses testing, best-subset regression, cross-validation, bootstrapping, Bayes method, and asymptotic methods that use maximum likelihood (i.e. the information-theoretic Akaike Information Criterion (AIC), the minimum-descriptive-length (MDL), and related minimum-message-length (MML) principle). The information-theoretic criteria attempt to quantify the tradeoff between model complexity and goodness-of-fit. (For an overview of these methods see Foster 2000).

The inclusion of a model selection criterion in the fitness function of an evolutionary algorithm (EA) is known as complexity-based fitness evaluation (Iba, 2000). Although this approach is largely unexplored, results to date have

been encouraging. Most experience with this approach has been with genetic programming (GP) to control decision-tree growth. Results of these studies have been promising (Iba, 2000 and references cited therein). However, very few studies have been done using complexity-based fitness evaluation (hereafter referred to as parsimony) in genetic algorithms (GA). Konagaya and Konoto (1993 as cited in Iba, 2000) used MDL for their fitness evaluation of a bioinformatics classification problem in order to minimize overlearning due to noise. Rolf *et al.* (1997) reported some preliminary results with model identification and parameter estimation of linear ARMA models using an EA that combined GA and evolutionary strategy (ES) operators. They tried different statistical criteria in their fitness function, and although the estimation of the error series by the EA was successful, accuracy of model selection was not very successful: identification of the correct model order was achieved only 20% of the time.

In contrast, Vesin and Grüter (1999) successfully used MDL in a simplex reproduction GA for selection of regressors in linear AR models and in nonlinear polynomial models. They were able to accurately identify the correct operating models, and they demonstrated fast convergence rates compared to exhaustive search techniques. Therefore, EA's with parsimony seem to offer the potential of an automated, efficient search technique for the best candidate models. In effect, the GA is orchestrating a competition among a community of candidate models while simultaneously optimizing parameter fit.

To date, the use of a GA with parsimony (GA+P) to select and fit more complex dynamic system models of the type often used in ecology has not been tried. In this paper we present results of experiments with a GA+P applied to a nested set of polynomials and dynamic simulation models. We introduce a modulo method of turning genes (model parameters) off and on, and compare it to letting selection and mutation drive non-contributing parameter values to zero. We also report on the effect of experimental error (Gaussian noise) in the data to the success and efficiency of model selection with GA+P.

2 EXPERIMENTAL APPROACH

Our approach was to choose an operating model from among a set of possible models, and to use that model to generate the “true” data (with or without added Gaussian noise) to which all of the competing models’ predictions were compared. First, we tested our approach on a set of polynomial models, and then we applied it to dynamic physiological-ecology models we built that simulate some biochemistry and biophysics that occur in a leaf undergoing photosynthesis. In both cases, we compared results with and without parsimony (GA+P and GA-P). Our overall objective was to evaluate the efficiency and success rate of GA+P in identifying the correct data-generating model, while accurately estimating the model parameters.

2.1 PARSIMONY

We implemented parsimony by the combination of an evolvable switch and AIC. The modulo (remainder) function was used to turn genes on and off. We used one of the less-significant digits of the gene value itself to determine whether the gene was active or not. By using modulus 2 on the integer-transformed gene value, we have, in effect, created a binary switch. Thus,

$$\begin{cases} \lfloor \text{Int}(a_i \times 10^k) \rfloor \pmod{2} = 0 \\ \text{or} \\ \lfloor \text{Int}(a_i \times 10^k) \rfloor \pmod{2} = 1 \end{cases}$$

for any i , where a_i refers to the gene value at position i on the chromosome, and k is of sufficient magnitude to shift a less-significant digit to the unit position of the resulting integer. Therefore, it is the value of the integer in the unit position that determines whether the gene is active or not. The number of active genes was then determined for calculating the penalty term of the AIC.

2.2 POLYNOMIAL MODELS

The general polynomial used to generate the set of candidate models for our preliminary tests was of the form

$$y(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

where $a_0, a_1, \dots, a_{n-1}, a_n$ are constants.

For these experiments, n was set to either five or nine. Initialization and evaluation by the GA effectively created the community of competing models comprised of subsets of the fifth- and ninth-order polynomials. Their genomes represent the coefficients of the various models. The operating model used for generating the “true” data for these experiments was a specific fourth-order model

$$y(x) = -20x^4 + 20x^2 + 14$$

evaluated over the domain $\{-1.0, -0.9, -0.8, \dots, 1.0\}$.

2.2.1 Polynomial Model Fitness

The total error for each candidate model was calculated as the log residual sum of squares (RSS)

$$e = \log \sum_{j=1}^m (y_j - \hat{y}_j)^2$$

where m is the number of data points, y_j is the true value of the correct operating model at point j and \hat{y}_j is the predicted value of a candidate model at point j . The \hat{y}_j value was calculated as

$$\hat{y}_j = \sum_{i=0}^n \{ \lfloor \text{Int}(a_i \times 10^8) \rfloor \pmod{2} \} a_i x_j^i$$

where n is the order of the complete polynomial model (five and nine for these experiments), and a_i refers to the gene (coefficient) value at position i on the chromosome. This estimated RSS was then transformed to the maximized log-likelihood (Burnham and Anderson 1998), and the AIC penalty term added. Therefore, the ensuing fitness of each candidate model was the true AIC.

2.3 DYNAMIC SIMULATION MODELS

The photosynthesis model contained six ODE’s that describe the carbon, water and heat budgets of the leaf over time. Soil water potential, herbivory, and ozone effects were also included in the model. External forcing functions accounted for the influence of light, temperature, humidity and wind velocity. Feedback loops linked the various model subcomponents together.

Thirteen genes represented model parameters associated with the stocks and fluxes of the carbon, water and heat budgets, and effects of ozone and herbivory. The “true” model output data were generated from a subset model whose genes for ozone and herbivore effects were turned off. These data comprised a parallel time series of ten metrics (photosynthetic rate, leaf carbon...) at 15-minute intervals over a 24-hour period.

2.3.1 Simulation Model Fitness

For each candidate simulation model, a sum of the relative error of each metric at each time point was calculated, and the penalty for the number of active parameters in the model was added to this sum. Therefore, fitness for these experiments was similar to a common analog of AIC (Hongzhi, 1989).

2.3.2 Software

An open-source parallel GA library, PGAPack, available from Argonne National Laboratory (Batavia, IL) was used for all experiments. The fitness functions and models were coded in C, optimized, and parallelized for SMP.

3 RESULTS

These experiments showed that a GA with parsimony could successfully select the correct data-generating model from among a community of candidate models (Figure 1). When noise was added to the data, both the polynomial and photosynthesis models without parsimony were largely unsuccessful in evolving the correct model. However, including parsimony in the fitness evaluation markedly improved the success rate, and the negative effect of noise was greatly reduced (Table 1).

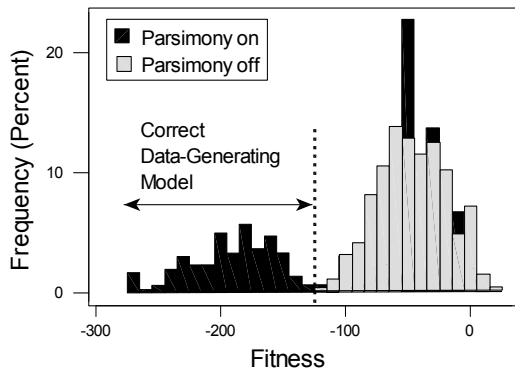


Figure 1: Success of the GA Evolving the Correct Data-Generating Model (Ninth-order Polynomial model with 300 Runs and No Noise).

Table 1: Effect of Parsimony (P) and Noise (N) on the Success of the GA Evolving the Correct Data-Generating Model. (Note: the numerator is the number of successes and the denominator is the total number of replicate runs)

Treatment →	- P	- P	+ P	+ P
↓ Model	- N	+ N	- N	+ N
Polynomial (5 th)	5/500	0/500	496/500	489/500
Polynomial (9 th)	0/300	0/300	121/300	78/300
Photosynthesis	9/30	3/30	29/30	29/30

The effect of overfitting is evident in Figure 2 where about 50% of the runs without parsimony achieved a better fit to the data by using additional parameters to fit the noise. Almost 98% of the runs with parsimony evolved the correct model and produced accurate and precise estimates of the parameters. The best GA+P runs with the polynomial models provided parameter estimates that were identical (to within 0.001) of those produced with a least-squares regression on the noisy data, after using the best-subset method with Mallows' Cp statistic for variable selection.

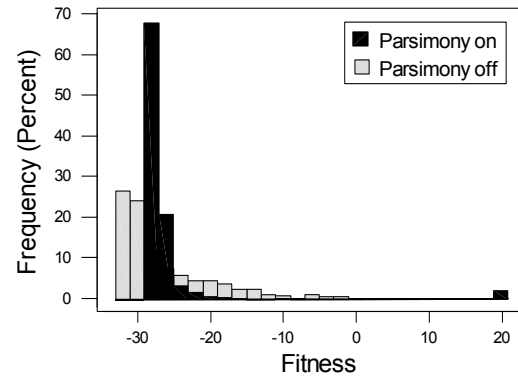


Figure 2: Effect of Noise on Overfitting (Fifth-order Polynomial model with 500 Runs and Noise).

4 CONCLUSIONS

Model selection with concurrent parameter fitting of ecological simulation models is feasible with this approach. AIC and the modulo method of regulating gene activity is an efficient way to implement parsimony. In particular, our modulo approach has a parallel in real organisms, where in some ribosomal RNA and tRNA genes, part of their coding sequence does double duty and serves as a regulatory switch for the gene.

Acknowledgments

USDA Hatch and DOE Computational Biology grants to the University of Vermont funded this work.

References

- Burnham, K. P. And D. R. Anderson (1998). *Model Selection and Inference: A Practical Information Theoretic Approach*. Springer-Verlag, New York.
- Foster, M.R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology* **44**(1):205-231.
- Hongzhi, A. (1989). Fast stepwise procedures of selection of variables by using AIC and BIC criteria. *Acta Mathematicae Applicatae Sinica* **5**:60-67
- Iba, H. (2000). Complexity-based fitness evaluation. In T. Bäck, D.B. Fogel and Z. Michalewicz (eds.) *Evolutionary Computation 2*, 15-24. Bristol, UK: IOP Publishing, Ltd.
- Rolf, S., Sprave, J. and W. Urfer (1997). Model identification and parameter estimation of ARMA models by means of evolutionary algorithms. Proceedings, IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, New York, IEEE Press, Piscataway, NJ, 237-243.
- Vesin, J-M. and R. Grüter (1999). Model selection using a simplex reproduction genetic algorithm. *Signal Processing* **78**:321-327.