00 GENETIC PROGRAMMING
Automatic Synthesis of Topologies and Numerical Parameters

John R. Koza

1. Introduction

Many mathematical algorithms are capable of solving problems by producing optimal (or near-optimal) numerical values for a prespecified set of parameters. However, for many practical problems, one cannot begin a search for the set of numerical values until one first ascertains the number of numerical values that one is seeking.

In fact, many practical problems of design and optimization entail first discovering an entire graphical structure (that is, a topology). After the topology is identified, optimal (or near-optimal) numerical values can be sought for the elements of the structure.

For example, if one is seeking an analog electrical circuit whose behavior satisfies certain prespecified high-level design goals, one must first ascertain the circuit's topology and then discover the numerical value of each electrical component in the circuit.

Specifically, the *topology* of an electrical circuit comprises

- the total number of electrical components, in the circuit,
- the type of each component (e.g., resistor, capacitor, transistor) at each location in the circuit, and
- a list of all the connections between the leads of the components.

The *sizing* of a circuit consists of the component value(s) for each component of the circuit that requires a component value.

The automatic synthesis of the topology and sizing of analog electrical circuits is a vexatious problem. As Aaserud and Nielsen (1995) noted

> "[M]ost … analog circuits are still handcrafted by the experts or so-called 'zahs' of analog design. The design process is characterized by a combination of experience and intuition and requires a thorough knowledge of the process characteristics and the detailed specifications of the actual product.

> "Analog circuit design is known to be a knowledge-intensive, multiphase, iterative task, which usually stretches over a significant period of time and is performed by designers with a large portfolio of skills. It is therefore considered by many to be a form of art rather than a science."

The purpose of a controller is to force, in a meritorious way, the actual response of a system (conventionally called the *plant*) to match a desired response (called the *reference signal*) (Dorf and Bishop 1998). Controllers are typically composed of signal processing blocks, such as integrators, differentiators, leads, lags, delays, gains, adders, inverters, subtractors, and multipliers.

A similarly vexatious situation arises if one is seeking the design of a controller whose behavior satisfies certain prespecified high-level design goals.

The topology of a controller comprises

- the total number of signal processing blocks in the controller,
- the type of each block (e.g., integrator, differentiator, lead, lag, delay, gain, adder, inverter, subtractor, and multiplier),
- the connections between the inputs and output of each block in the controller and the external input and external output points of the controller.

The *tuning* (sizing) of a controller consists of the parameter values associated with each signal processing block.

A parallel situation arises in connection with networks of chemical reactions (metabolic pathways).

The concentrations of substrates, products, and intermediate substances participating in a network of chemical reactions are modeled by non-linear continuous-time differential equations, including various first-order rate laws, second-order rate laws, power laws, and the Michaelis-Menten equations (Voit 2000). The concentrations of catalysts (e.g., enzymes) control the rates of many chemical reactions in living things.

The *topology* of a network of chemical reactions comprises

- the total number of reactions,
- the number of substrates consumed by each reaction,
- the number of products produced by each reaction,
- the pathways supplying the substrates (either from external sources or other reactions in the network) to each reaction,
- the pathways dispersing each reaction's products (either to other reactions or external outputs), and
- an identification of whether a particular enzyme acts as a catalyst.

The *sizing* for a network of chemical reactions consists of all the numerical values associated with the network (e.g., the rates of each reaction).

A similarly vexatious situation arises if one is seeking the design of an antenna whose behavior satisfies certain prespecified high-level design goals.

While it might seem difficult or impossible to automatically create both the topology and numerical parameters for a complex structure merely from a high-level statement of the structure's design goals, recent work has demonstrated that genetic programming can automatically create complex structures that exhibit prespecified behavior in fields where the structure's behavior is modeled by differential equations (both linear and non-linear) or by other equations (e.g., Maxwell's equations).

In this chapter, we will demonstrate that a biologically motivated algorithm (genetic programming) can automatically synthesize both a graphical structure (the topology) and a set of optimal or near-optimal numerical values for each element of

- analog electrical circuits (section 3),
- controllers (section 4),
- antennas (section 5), and
- networks of chemical reactions (metabolic pathways) (section 6).

2. Genetic Programming

Genetic programming progressively breeds a population of computer programs over a series of generations by starting with a primordial ooze of thousands of randomly created computer programs and using the Darwinian principle of natural selection, recombination (crossover), mutation, gene duplication, gene deletion, and certain mechanisms of developmental biology.

Genetic programming breeds computer programs to solve problems by executing the following three steps:

(1) Generate an initial population of compositions (typically random) of the functions and terminals of the problem.

(2) Iteratively perform the following substeps (referred to herein as a generation) on the population of programs until the termination criterion has been satisfied:

(A) Execute each program in the population and assign it a fitness value using the fitness measure.

(B) Create a new population of programs by applying the following operations. The operations are applied to program(s) selected from the population with a probability based on fitness (with reselection allowed).

(i) Reproduction: Copy the selected program to the new population.

(ii) Crossover: Create a new offspring program for the new population by recombining randomly chosen parts of two selected programs.

(iii) Mutation: Create one new offspring program for the new population by randomly mutating a randomly chosen part of the selected program.

(iv) Architecture-altering operations: Select an architecture-altering operation from the available repertoire of such operations and create one new offspring program for the new population by applying the selected architecture-altering operation to the selected program.

(3) Designate the individual program that is identified by result designation (e.g., the best-so-far individual) as the result of the run of genetic programming. This result may be a solution (or an approximate solution) to the problem.

Genetic programming is described in the book *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (Koza 1992; Koza and Rice 1992), the book *Genetic Programming II: Automatic Discovery of Reusable Programs* (Koza 1994a, 1994b), and the book *Genetic Programming III: Darwinian Invention and Problem Solving* (Koza, Bennett, Andre, and Keane 1999; Koza, Bennett, Andre, Keane, and Brave 1999).

Genetic programming is an extension of the genetic algorithm (Holland 1975) in which the population being bred consists of computer programs.

Genetic programming starts with an initial population of randomly generated computer programs composed of the given primitive functions and terminals. The programs in the population are, in general, of different sizes and shapes. The creation of the initial random population is a blind random search of the space of computer programs composed of the problem's available functions and terminals.

On each generation of a run of genetic programming, each individual in the population of programs is evaluated as to its fitness in solving the problem at hand. The programs in generation 0 of a run almost always have exceedingly poor fitness for non-trivial problems of interest. Nonetheless, some individuals in a population will turn out to be somewhat more fit than others. These differences in performance are then exploited so as to direct the search into promising areas of the search space. The Darwinian principle of reproduction and survival of the fittest is used to probabilistically select, on the basis of fitness, individuals from the population to participate in various operations. A small percentage (e.g., 9%) of the selected individuals are reproduced (copied) from one generation to the next. A very small percentage (e.g. 1%) of the selected individuals are mutated in a random way. Mutation can be viewed as an undirected local search mechanism. The vast majority of the selected individuals (e.g., 90%) participate in the genetic operation of crossover (sexual recombination) in which two offspring programs are created by recombining genetic material from two parents.

The creation of the initial random population and the creation of offspring by the genetic operations are all performed so as to create syntactically valid, executable programs. After the genetic operations are performed on the current generation of the population, the population of offspring (i.e., the new generation) replaces the old generation. The tasks of measuring fitness, Darwinian selection, and genetic operations are then iteratively repeated over many generations. The computer program resulting from this simulated evolutionary process can be the solution to a given problem or a sequence of instructions for constructing the solution.

Probabilistic steps are pervasive in genetic programming. Probability is involved in the creation the individuals in the initial population, the selection of individuals to participate in the genetic operations (e.g., reproduction, crossover, and mutation), and the selection of crossover and mutation points within parental programs.

The dynamic variability of the size and shape of the computer programs that are created during the run is an important feature of genetic programming. It is often difficult and unnatural to try to specify or restrict the size and shape of the eventual solution in advance.

The individual programs that are evolved by genetic programming are typically multi-branch programs consisting of one or more result-producing branches and zero, one, or more automatically defined functions (subroutines).

The *architecture* of such a multi-branch program involves

(1) the total number of automatically defined functions,

(2) the number of arguments (if any) possessed by each automatically defined function, and

(3) if there is more than one automatically defined function in a program, the nature of the hierarchical references (including recursive references), if any, allowed among the automatically defined functions.

Architecture-altering operations enable genetic programming to automatically determine the number of automatically defined functions, the number of arguments that each possesses, and the nature of the hierarchical references, if any, among such automatically defined functions.

Additional information on genetic programming can be found in books such as Banzhaf, Nordin, Keller, and Francone 1998; books in the series on genetic programming from Kluwer Academic Publishers such as Langdon 1998, Ryan 1999, and Wong and Leung 2000; in edited collections of papers such as the *Advances in Genetic Programming* series of books from the MIT Press (Kinnear 1994; Angeline and Kinnear 1996; Spector, Langdon, O'Reilly, and Angeline 1999); in the proceedings of the Genetic Programming Conference held between 1996 and 1998 (Koza, Goldberg, Fogel, and Riolo 1996; Koza, Deb, Dorigo, Fogel, Garzon, Iba, and Riolo 1997; Koza, Banzhaf, Chellapilla, Deb, Dorigo, Fogel, Garzon, Goldberg, Iba, and Riolo 1998); in the proceedings of the annual Genetic and Evolutionary Computation Conference (combining the annual Genetic Programming Conference and the International Conference on Genetic Algorithms) held starting in 1999 (Banzhaf, Daida, Eiben, Garzon, Honavar, Jakiela, and Smith, 1999; Whitley, Goldberg, Cantu-Paz, Spector, Parmee, and Beyer 2000); in the proceedings of the annual Euro-GP conferences held starting in 1998 (Banzhaf, Poli, Schoenauer, and Fogarty 1998; Poli, Nordin, Langdon, and Fogarty 1999; Poli, Banzhaf, Langdon, Miller, Nordin, and Fogarty 2000); at web sites such as `www.genetic-programming.org`; and in the Genetic Programming and Evolvable Machines journal (from Kluwer Academic Publishers).

3. Automatic Synthesis of Analog Electrical Circuits

Genetic programming is capable of automatically creating both the topology and sizing (component values) for analog electrical circuits composed of transistors, capacitors, resistors, and other components merely by specifying the circuit's behavior.

This automatic synthesis of circuits is performed by genetic programming even though there is no general mathematical method (prior to genetic programming) for creating (synthesizing) both the topology and sizing (component values) of analog electrical circuits from the circuit's desired (or observed) behavior (Aaserud and Nielsen 1995; Koza, Bennett, Andre, and Keane 1999).

For purposes of illustration, we discuss

- a lowpass filter circuits using a fitness measure based on the frequency-domain behavior of circuits, and
- a computational circuit employing transistors and using a fitness measure based on the time-domain behavior of circuits.

3.1.1. Lowpass Filter Circuit

A simple *filter* is a one-input, one-output electronic circuit that receives a signal as its input and passes the frequency components of the incoming signal that lie in a specified range (called the *passband*) while suppressing the frequency components that lie in all other frequency ranges (the *stopband*).

A *lowpass filter* passes all frequencies below a certain specified frequency, but stops all higher frequencies.

Figure FFF1 shows the frequency domain behavior of an illustrative lowpass filter in which the boundary of the passband is at 1,000 Hz and the boundary of the stopband is 2,000 Hz. The horizontal axis represents the frequency of the incoming signal and ranges over five decades of frequencies between 1 Hz and 100,000 Hz on a logarithmic scale. The vertical axis represents the peak voltage of the output and ranges between 0 to 1 Volts on a linear scale. This figure shows that when the input to the circuit consists of a sinusoidal signal with any frequency from 1 Hz to 1,000 Hz, the output is a sinusoidal signal with an amplitude of a full 1 Volt. This figure also shows that when the input to the circuit consists of a sinusoidal signal with any frequency from 2,000 Hz to 100,000 Hz, the amplitude of the output is essentially 0 Volts. The region between 1,000 Hz and 2,000 Hz is a transition region where the voltage varies between 1 Volts (at 1,000 Hz) and essentially 0 Volts (at 2,000 Hz).

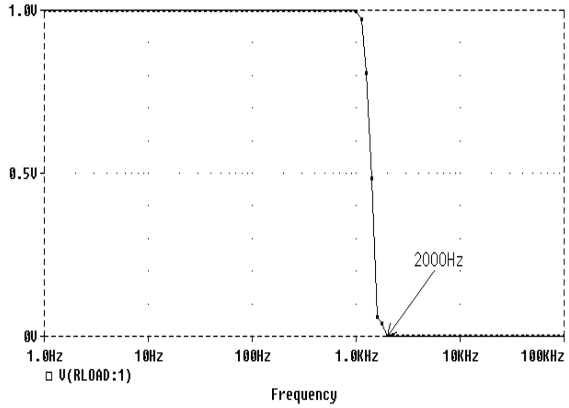

Figure FFF1 Frequency domain behavior of a lowpass filter.

NOTE TO EDITORS: The FFF's and BBB's will be deleted in the final version.

Genetic programming is capable of automatically creating both the topology and sizing (component values) for lowpass filters (and other filter circuits, such as highpass filters, bandpass filters, bandstop filters, and filters with multiple passbands and stopbands).

A filter circuit may be evolved using a fitness measure based on frequency domain behavior. In particular, the fitness of an individual circuit is the sum, over 101 values of frequency between 1 Hz and 100,000 Hz (equally space on a logarithmic scale), of the absolute value of the difference between the individual circuit's output voltage and the ideal voltage for an ideal lowpass filter for that frequency (i.e., the voltages shown in figure FFF1).

For example, one run of genetic programming synthesized the lowpass filter circuit of figure FFF2.

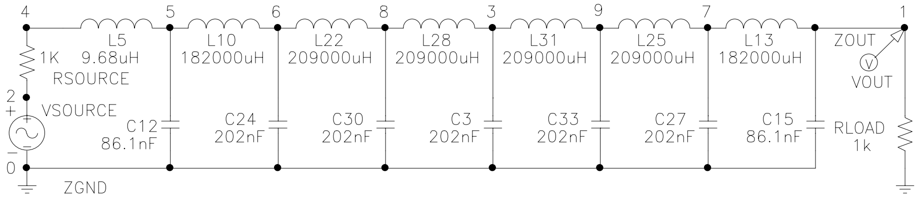

Figure FFF2 Lowpass filter created by genetic programming that infringes on Campbell's patent.

The evolved circuit of figure FFF2 is what is now called a cascade (ladder) of identical π sections (Koza, Bennett, Andre, and Keane 1999, chapter 25). The evolved circuit has the recognizable topology of the circuit for which George Campbell of American Telephone and Telegraph received U. S. patent 1,227,113 in 1917. Claim 2 of Campbell’s patent covered,

> “An electric wave filter consisting of a connecting line of negligible attenuation composed of a plurality of sections, each section including a capacity element and an inductance element, one of said elements of each section being in series with the line and the other in shunt across the line, said capacity and inductance elements having precomputed values dependent upon the upper limiting frequency and the lower limiting frequency of a range of frequencies it is desired to transmit without attenuation, the values of said capacity and inductance elements being so proportioned that the structure transmits with practically negligible attenuation sinusoidal currents of all frequencies lying between said two limiting frequencies, while attenuating and approximately extinguishing currents of neighboring frequencies lying outside of said limiting frequencies.”

In addition to possessing the topology of the Campbell filter, the numerical values of all the components in the evolved circuit closely approximate the numerical values taught in Campbell’s 1917 patent.

Another run of genetic programming synthesized the lowpass filter circuit of figure FFF3. As before, this circuit was evolved using the previously described fitness measure based on frequency domain behavior.

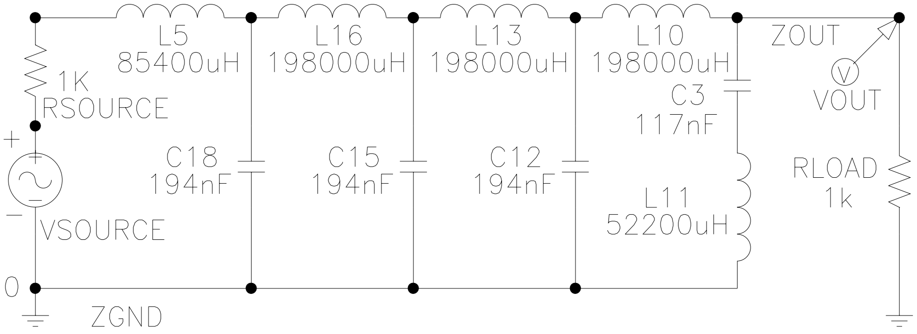

Figure FFF3 Lowpass filter created by genetic programming that infringes on Zobel's patent.

This evolved circuit differs from the Campbell filter in that its final section consists of both a capacitor and inductor. This filter is an improvement over the Campbell filter because its final section confers certain performance advantages on the circuit. This circuit is equivalent to what is called a cascade of three symmetric T-sections and an *M*-derived half section (Koza, Bennett, Andre, and Keane 1999, chapter 25). Otto Zobel of American Telephone and Telegraph Company invented and received a patent for an “*M*-derived half section” used in conjunction with one or more “constant K” sections. Again, the numerical values of all the components in this evolved circuit closely approximate the numerical values taught in Zobel's 1925 patent.

Seven circuits created using genetic programming infringe on previously issued patents ((Koza, Bennett, Andre, and Keane 1999). Others duplicate the functionality of previously patented inventions in novel ways.

In both of the foregoing examples, genetic programming automatically created both the topology and sizing (component values) of the entire filter circuit by using a fitness measure expressed in terms of the signal observed at the single output point (the probe point labeled VOUT in the figures).

3.1.2. Squaring Computational Circuit

Because filters discriminate on incoming signals based on frequency, the lowpass filter circuit was automatically synthesized using a fitness measure based on the behavior of the circuit in the frequency domain. However, for many circuits, it is appropriate to synthesize the circuit using a fitness measure based on the behavior of the circuit in the time domain.

An analog electrical circuit whose output is a well-known mathematical function (e.g., square, square root) is called a *computational circuit*.

Figure FFF4 shows a squaring circuit composed of transistors, capacitors, and resistors that was automatically synthesized using a fitness measure based on the behavior of the circuit in the time domain (Mydlowec and Koza 2000).

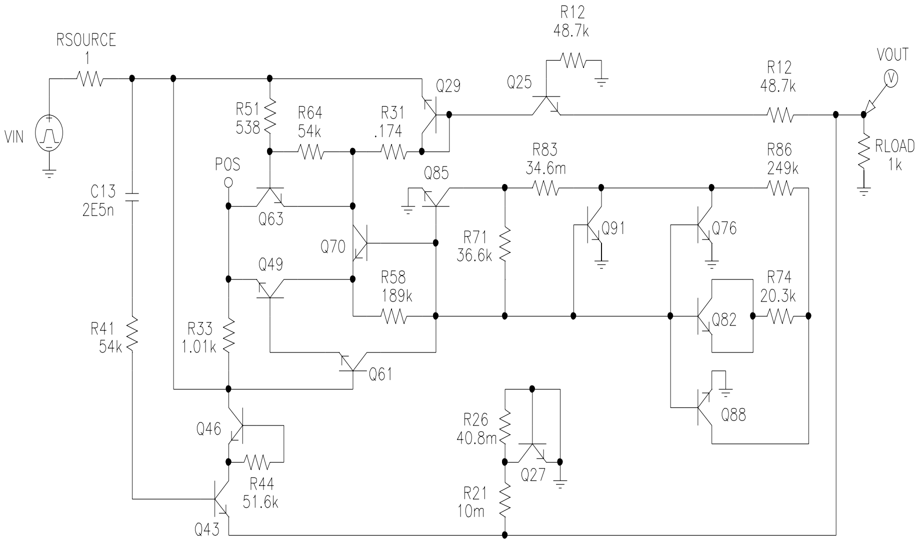

Figure FFF4 Squaring circuit created by genetic programming.

This circuit was evolved using a fitness measure based on time-varying input signals. In particular, fitness was the sum, taken at certain sampled times for four different time-varying input signals, of the absolute value of the difference between the individual circuit's output voltage and the desired output voltage (i.e., the square of the voltage of the input signal at the particular sampled time).

The four input signals were structured to provide a representative mixture of input values. All of the input signals produce outputs that are well within the range of voltages that can be handled by transistors (i.e., below 4 volts). For example, one of the input signals is a rising ramp whose value remains at 0 up to 0.2 seconds and then rises to 2 Volts between 0.2 seconds and 1.0 seconds. Figure FFF5 shows the output voltage produced by the evolved circuit for the rising ramp input superimposed on the (virtually indistinguishable) correct output voltage for the squaring function. As can be seen, as soon as the input signal becomes non-zero, the output is a parabolic-shaped curve representing the square of the incoming voltage.

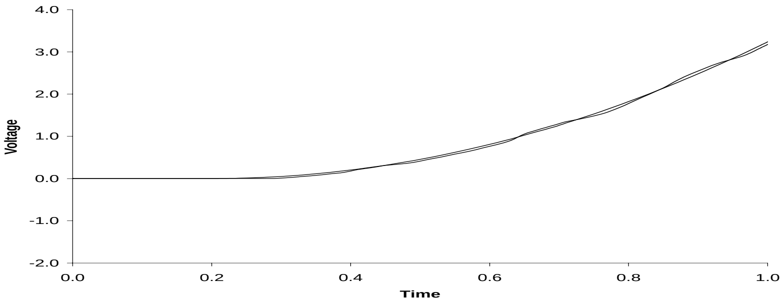

Figure FFF5 Output for rising ramp input for squaring circuit.

4. Automatic Synthesis of Controllers

Genetic programming is capable of automatically creating both the topology and sizing (tuning) for controllers composed of time-domain blocks (such as integrators, differentiators, multipliers, adders, delays, gains, leads, and lags) merely by specifying the controller's effect on the to-be-controlled plant (Keane, Yu, and Koza 2000; Koza, Keane, Yu, Bennett, Mydlowec, and Stiffelman 1999; Koza, Keane, Bennett, Yu, Mydlowec, and Stiffelman, Oscar 1999; Koza, Keane, Yu, Bennett, and Mydlowec 2000; Koza, Keane, Yu, Mydlowec, and Bennett 2000a, 2000b; Koza, Yu, Keane, and Mydlowec 2000; Yu, Keane, and Koza. 2000). This automatic synthesis of controllers from data is performed by genetic programming even though there is no general mathematical method for creating both the topology and sizing for controllers from a high-level statement of the design goals for the controller.

In the PID type of controller, the controller's output is the sum of proportional (P), integrative (I), and derivative (D) terms based on the difference between the plant's output and the reference signal. Albert Callender and Allan Stevenson of Imperial Chemical Limited of Northwich, England received U.S. Patent 2,175,985 in 1939 for the PI and PID controller.

Claim 1 of Callender and Stevenson (1939) covers what is now called the PI controller,

> "A system for the automatic control of a variable characteristic comprising means proportionally responsive to deviations of the characteristic from a desired value, compensating means for adjusting the value of the characteristic, and electrical means associated with and actuated by responsive variations in said responsive means, for operating the compensating means to correct such deviations in

> conformity with the sum of the extent of the deviation and the summation of the deviation."

Claim 3 of Callender and Stevenson (1939) covers what is now called the PID controller,

> "A system as set forth in claim 1 in which said operation is additionally controlled in conformity with the rate of such deviation."

The vast majority of automatic controllers used by industry are of the PI or PID type. However, it is generally recognized by leading practitioners in the field of control that PI and PID controllers are not ideal (Astrom and Hagglund 1995; Boyd and Barratt 1991).

There is no preexisting general-purpose analytic method (prior to genetic programming) for automatically creating both the topology and tuning of a controller for arbitrary linear and non-linear plants that can simultaneously optimize prespecified performance metrics. The performance metrics used in the field of control include, among others,

- minimizing the time required to bring the plant output to the desired value (as measured by, say, the integral of the time-weighted absolute error),
- satisfying time-domain constraints (involving, say, overshoot and disturbance rejection),
- satisfying frequency domain constraints (e.g., bandwidth), and
- satisfying additional constraints, such as limiting the magnitude of the control variable or the plant's internal state variables.

We employ a problem involving control of a two-lag plant (described by Dorf and Bishop 1998, page 707) to illustrate the automatic synthesis of controllers by means of genetic programming. The problem entails synthesizing the design of both the topology and parameter values for a controller for a two-lag plant such that plant output reaches the level of the reference signal so as to minimize the integral of the time-weighted absolute error, such that the overshoot in response to a step input is less than 2%, and such that the controller is robust in the face of significant variation in the plant's internal gain, K, and the plant's time constant, τ.

Genetic programming routinely creates PI and PID controllers infringing on the 1942 of Callender and Stevenson patent during intermediate generations of runs of genetic programming on controller problems. However, the PID controller is not the best possible controller for this (and many) problems.

Figure FFF6 shows the block diagram for the best-of-run controller evolved during one run of this problem. In this figure, $R(s)$ is the reference signal; $Y(s)$ is the plant output; and $U(s)$ is the controller's output (control variable). This evolved controller is 2.42 times better than the Dorf and Bishop (1998) controller as measured by the criterion used by Dorf and Bishop. In addition, this evolved controller has only 56% of the rise time in response to the reference input, has only 32% of the settling time, and is 8.97 times better in terms of suppressing the effects of disturbance at the plant input.

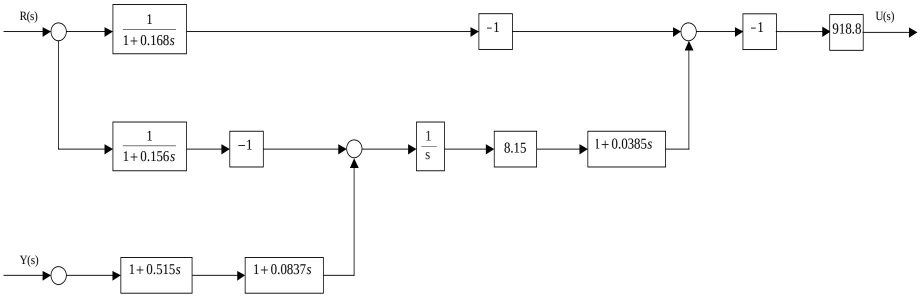

Figure FFF6 Evolved controller that infringes on Jones' patent.

This genetically evolved controller differs from a conventional PID controller in that it employs a second derivative processing block. Specifically, after applying standard manipulations to the block diagram of this evolved controller, the transfer function for the best-of-run controller can be expressed as a transfer function for a pre-filter and a transfer function for a compensator. The transfer function for the pre-filter, $G_{p32}(s)$, for the best-of-run individual from generation 32 is

$$G_{p32}(s) = \frac{1(1+.1262s)(1+.2029s)}{(1+.03851s)(1+.05146)(1+.08375)(1+.1561s)(1+.1680s)}$$

The transfer function for the compensator, $G_{c32}(s)$, is

$$G_{c32}(s) = \frac{7487(1+.03851s)(1+.05146s)(1+.08375s)}{s} = \frac{7487.05+1300.63s+71.2511s^2+1.2426s^3}{s}$$

The s^3 term (in conjunction with the s in the denominator) indicates a second derivative. Thus, the compensator consists of a second derivative in addition to proportional, integrative, and derivative functions. As it happens, Harry Jones of The Brown Instrument Company of Philadelphia received U. S. Patent 2,282,726 for this kind of controller topology in 1942.

Claim 38 of the Jones patent (Jones 1942) states,

> "In a control system, an electrical network, means to adjust said network in response to changes in a variable condition to be controlled, control means responsive to network adjustments to control said condition, reset means including a reactance in said network adapted following an adjustment of said network by said first means to initiate an additional

> network adjustment in the same sense, and rate control means included in said network adapted to control the effect of the first mentioned adjustment in accordance with the second or higher derivative of the magnitude of the condition with respect to time."

Note that the human user of genetic programming did not preordain, prior to the run (i.e., as part of the preparatory steps for genetic programming), that a second derivative should be used in the controller (or, from that matter, even that a P, I, or D block should be used). Genetic programming automatically discovered that the second derivative element (along with the P, I, and D elements) were useful in producing a good controller for this particular problem. That is, necessity was the mother of invention.

Similarly, the human who initiated this run of genetic programming did not preordain any particular topological arrangement of proportional, integrative, derivative, second derivative, or other functions within the automatically created controller. Instead, genetic programming automatically created a controller for the given plant without the benefit of user-supplied information concerning the total number of processing blocks to be employed in the controller, the type of each processing block, the topological interconnections between the blocks, the values of parameters for the blocks, or the existence of internal feedback (none in this instance) within the controller.

5. Automatic Synthesis of Antennas

An antenna is a device for receiving or transmitting electromagnetic waves. An antenna may receive an electromagnetic wave and transform it into a signal on a transmission line. Alternately, an antenna may transform a signal from a transmission line into an electromagnetic wave that is then propagated in free space.

Maxwell's equations govern the electromagnetic waves generated and received by antennas. The behavior and characteristics of many antennas can be determined by simulation. For example, the *Numerical Electromagnetics Code* (NEC) is a method-of-moments (MoM) simulator for wire antennas that was developed at the Lawrence Livermore National Laboratory (Burke 1992).

The task of analyzing the characteristics of a given antenna is difficult. The task of synthesizing the design of an antenna with specified characteristics typically calls for considerable creativity on the part of the antenna engineer (Balanis 1982; Stutzman and Thiele 1998; Linden 1997).

Genetic programming is capable of discovering both the topological and numerical aspects of a satisfactory antenna design from a high-level specification of the antenna's behavior. In one particular problem (Comisky, Yu, and Koza 2000), genetic programming automatically discovered the design for a satisfactory antenna composed of wires for maximizing gain in a preferred direction over a specified range of frequencies, having a reasonable value of voltage standing wave ratio when the antenna is fed by a transmission line with a specified characteristic impedance, and fitting into a specified bounding rectangle. The design that genetic programming discovered included

(1) the number of directors in the antenna,

(2) the number of reflectors,

(3) the fact that the driven element, the directors, and the reflector are all single straight wires,

(4) the fact that the driven element, the directors, and the reflector are all arranged in parallel,

(5) the fact that the energy source (via the transmission line) is connected only to the driven element — that is, the directors and reflectors are parasitically coupled.

The last three of the above characteristics discovered by genetic programming are the defining characteristics of an inventive design conceived in the early years of the field of antenna design (Uda 1926, 1927; Yagi 1928). Figure FFF7 shows the antenna created by genetic programming. It is an example of what is now called a Yagi-Uda antenna. It is approximately the same length as the conventional Yagi-Uda antenna that a human designer might develop in order to satisfy this problem's requirements (concerning gain).

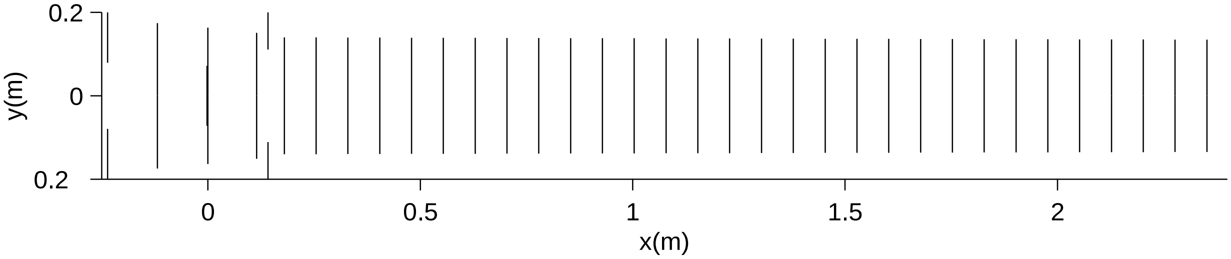

Figure FFF7 Antenna design created by genetic programming.

6. Automatic Synthesis of Metabolic Pathways

A living cell can be viewed as a dynamical system in which a large number of different substances react continuously and non-linearly with one another. In order to understand the behavior of a continuous non-linear dynamical system with numerous interacting parts, it is usually insufficient to study behavior of each part in isolation. Instead, the behavior must usually be analyzed as a whole (Tomita, Hashimoto, Takahashi, Shimizu, Matsuzaki, Miyoshi, Saito, Tanida, Yugi, Venter, and Hutchison 1999; Voit 2000).

Biochemists and others have historically determined the topology and sizing of networks of chemical reactions, such as metabolic pathways, through meticulous study of particular networks of interest. However, vast amounts of time-domain data are now becoming available concerning the concentration of biologically important chemicals in living organisms (McAdams and Shapiro 1995; Loomis and Sternberg 1995; Arkin, Shen, and Ross 1997; Yuh, Bolouri, and Davidson 1998; Laing, Fuhrman, and Somogyi 1998; D'haeseleer, Wen, Fuhrman, and Somogyi 1999). Such data include both gene expression data (obtained from microarrays) and data on the concentration of substances participating in metabolic pathways.

The question arises as to whether it is possible to start with observed time-domain concentrations of final product substance(s) and automatically create both the topology of the network of chemical reactions and the sizing of the network. In other words, is it possible to automate the process of reverse engineering a network of chemical reactions from data?

Intuitively, it might seem difficult or impossible to automatically infer both the topology and numerical parameters for a complex network from observed data. However, such intuition may be misleading.

Our approach to the problem of automatically creating both the topology and sizing of a network of chemical reactions involves

(1) establishing a representation for chemical networks involving symbolic expressions (S-expressions) and program trees that are composed of functions and terminals and that can be progressively bred (and improved) by genetic programming,

(2) converting each individual program tree in the population into an analog electrical circuit representing a network of chemical reactions,

(3) obtaining the behavior of the individual network of chemical reactions by simulating the corresponding electrical circuit,

(4) defining a fitness measure that measures how well the behavior of an individual network matches the observed time-domain data concerning concentrations of product substances, and

(5) using the fitness measure to enable genetic programming to breed a population of improving program trees.

6.1. Phospholipid Cycle

The best-of-run individual (figure FFF8) appears in generation 225. Its fitness is almost zero (0.054). This closely matches the observed data for all data points. In addition to having the same topology as the correct metabolic pathway, the rate constants of three of the four reactions of this network match the correct rates (to three significant digits). The fourth rate is within less than 2% of the correct rate (i.e., the rate of EC 3.1.3.21 is 1.17 compared with 1.19 for the correct network).

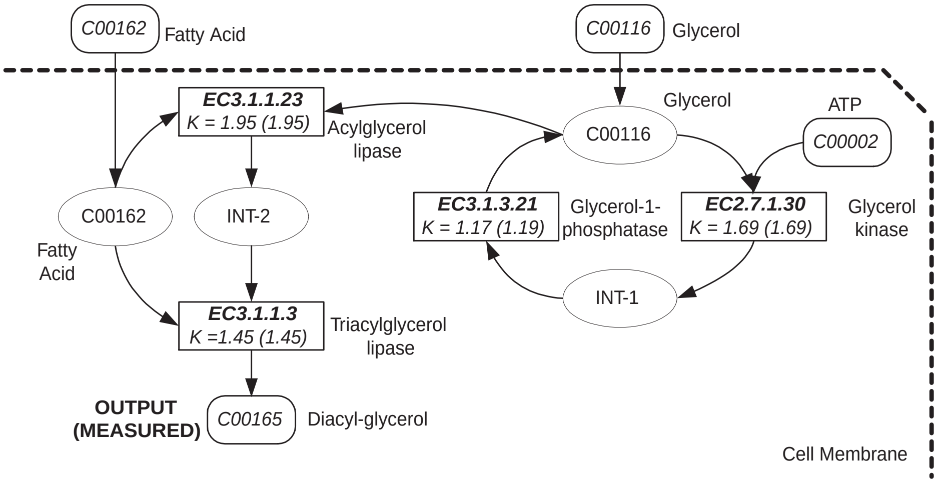

Figure FFF8 Network of chemical reactions for the best-of-run individual from generation 225.

Figure FFF9 shows the electrical circuit for the best-of-run individual from generation 225.

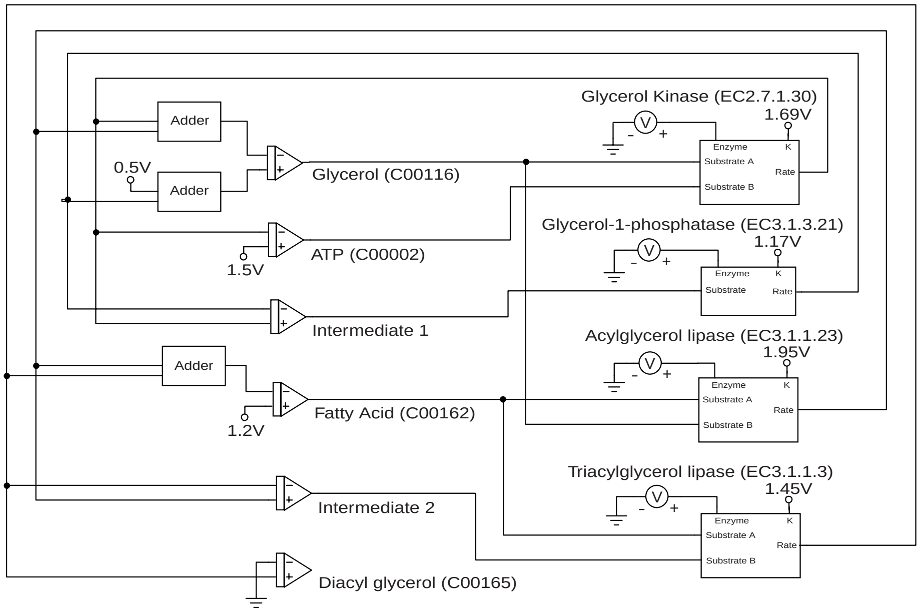

Figure FFF9 Electrical circuit for the best-of-run individual from generation 225.

In the best-of-run network, the rate of the two-substrate, one-product reaction catalyzed by Triacylglycerol lipase (EC 3.1.1.3) (found at the very bottom of figures FFF8 and FFF9) that produces the final product diacyl-glycerol (C00165) is given by

$$\frac{d[C00165]}{dt} = 1.45[C00162][INT\_2][\mathrm{EC}\,3.1.1.3]\,.$$

Note that genetic programming has correctly determined that the reaction that produces the network's final product diacyl-glycerol (C00165) has two substrates and one product; it has correctly identified enzyme EC3.1.1.3 as the catalyst for this final reaction; it has correctly determined the rate of this final reaction as 1.45; and it has correctly identified the externally supplied substance, fatty acid (C00162), as one of the two substrates for this final reaction.

Of course, genetic programming has no way of knowing that biochemists call the intermediate substance (INT_2) by the name Monoacyl-glycerol (C01885). It has, however, correctly determined that an intermediate substance is needed as one of the two substrates of the network's final reaction and that this intermediate substance should, in turn, be produced by a particular other reaction (described next).

The rate of the two-substrate, one-product reaction catalyzed by Acylglycerol lipase (EC3.1.1.23) that produces intermediate substance INT_2 is

$$\frac{d[INT\_2]}{dt} = 1.95[C00162][C00116][\mathrm{EC}\,3.1.1.23]\text{ - }1.45[C00162][INT\_2][\mathrm{EC}\,3.1.1.3]\,.$$

Again, genetic programming has correctly determined that the reaction that produces the intermediate substance (INT_2) has two substrates and one product; it has correctly identified enzyme EC3.1.1.23 as the catalyst for this reaction; it has correctly determined the rate of this reaction as 1.95; it has correctly identified two externally supplied substance, fatty acid (C00162) and glycerol (C00116), as the two substrates for this reaction.

The rate of the two-substrate, one-product reaction catalyzed by Glycerol kinase (EC2.7.1.30) that produces intermediate substance INT_1 in the internal loop is

$$\frac{d[INT\_1]}{dt} = 1.69[C00116][C00002][\mathrm{EC}\,2.7.1.30]\text{ - }1.17[INT\_1][\mathrm{EC}\,3.1.3.21]\,.$$

Note that the numerical rate constant of 1.17 in the above equation is within less than 2% of the correct rate of 1.19..

Here again, genetic programming has correctly determined that the reaction that produces the intermediate substance (INT_1) has two substrates and one product; it has correctly identified enzyme EC2.7.1.30 as the catalyst for this reaction; it has almost correctly determined the rate of this reaction to be 1.17 (whereas the correct rate is 1.19); it has correctly identified two externally supplied substance, glycerol (C00116) and the cofactor ATP (C00002), as the two substrates for this reaction.

Genetic programming has no way of knowing that biochemists call the intermediate substance (INT_1) by the name sn-Glycerol-3-Phosphate (C00093). Genetic programming has, however, correctly determined that an intermediate substance is needed as the single substrate of the reaction catalyzed by Glycerol-1-phosphatase (EC3.1.3.21) and that this intermediate substance should, in turn, be produced by the reaction catalyzed by Glycerol kinase (EC2.7.1.30).

The rate of supply and consumption of cofactor ATP (C00002) is

$$\frac{d[ATP]}{dt} = 1.5 - 1.69[C00116][C00002][\mathrm{EC}\,2.7.1.30]$$

The rate of supply and consumption of fatty acid (C00162) is

$$\frac{d[C00162]}{dt} = 1.2 - 1.95[C00162][C00116][\mathrm{EC}\,3.1.1.23] \text{ - } 1.45[C00162][INT\_2][\mathrm{EC}\,3.1.1.3]\,.$$

The rate of supply, consumption, and production of glycerol (C00116) is

$$\frac{d[C00116]}{dt} = 0.5 + 1.17[INT\_1][\mathrm{EC}\,3.1.3.21] \text{ - } 1.69[C00116][C00002][\mathrm{EC}\,2.7.1.30] \text{ - } 1.95[C00162][C00116][\mathrm{EC}\,3.1.1.23]$$

Again, note that the numerical rate constant of 1.17 in the above equation is slightly different from the correct rate.

Notice the internal feedback loop in which C00116 is both consumed and produced.

In summary, driven only by the time-domain concentration values of the final product C00165 (diacyl-glycerol), genetic programming created the entire metabolic pathway, including

- topological features such as the internal feedback loop,
- topological features such as a bifurcation point where one substance is distributed to two different reactions,
- topological features such as an accumulation point where one substance is accumulated from two sources, and
- numerical rates (sizing) for all reactions.

Notice that genetic programming created the entire metabolic pathway, including topological features (such as the internal feedback loop, the bifurcation point, and the accumulation point) and all numerical rate parameter values (sizing) of the metabolic pathway. Genetic programming also determined that two intermediate substances (INT_1 and INT_2) would be used. Genetic programming did this using only the time-domain concentration values of the final product C00165 (diacyl-glycerol,).

Both the topology and sizing of the metabolic pathway were created by using 270 time-domain values of the final product. This example (and the one below) demonstrate the principle that it is possible to reverse engineer a metabolic pathway from observed data concerning the concentration values of its final output.

For additional details, see Koza, Mydlowec, Lanza, Yu, and Keane 2000.

6.2. Synthesis and Degradation of Ketone Bodies

We proceed in the same way to automatically create a metabolic pathway for the synthesis and degradation of ketone bodies.

The best-of-run network appears in generation 97 (figure FFF10). It has a fitness of 0.000 and scores 270 hits. This individual has the same topology as the correct metabolic pathway and the same rates (to three significant digits) for each of the three reactions.

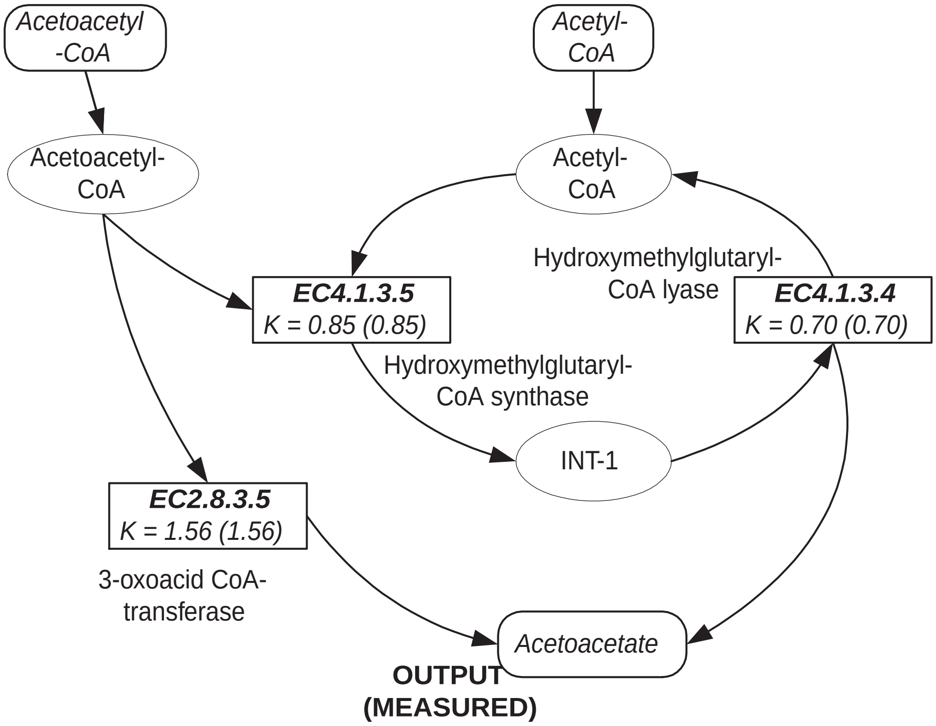

Figure FFF10 Best network of generation 97

In summary, driven only by the time-domain concentration values of Acetoacetate (the final product), the evolutionary process of genetic programming created the entire metabolic pathway, including

- topological features such as the internal feedback loop,
- topological features such as a bifurcation point where one substance is distributed to two different reactions,
- topological features such as an accumulation point where one substance is accumulated from two sources, and
- numerical rates (sizing) for all reactions.

7. Conclusions

This chapter has demonstrated that a biologically motivated algorithm (genetic programming) is capable of automatically synthesizing both the topology of complex graphical structures and optimal or near-optimal numerical values for all elements of the structure possessing parameters.

References

Aaserud, O. and Nielsen, I. Ring. 1995. Trends in current analog design: A panel debate. *Analog Integrated Circuits and Signal Processing*. 7(1) 5-9.

Angeline, Peter J. and Kinnear, Kenneth E. Jr. (editors). 1996. *Advances in Genetic Programming 2*. Cambridge, MA: The MIT Press.

Arkin, Adam, Shen, Peidong, and Ross, John. 1997. A test case of correlation metric construction of a reaction pathway from measurements. *Science*. 277. Pages 1275 - 1279. August 29, 1997.

Astrom, Karl J. and Hagglund, Tore. 1995. *PID Controllers: Theory, Design, and Tuning.* Second Edition. Research Triangle Park, NC: Instrument Society of America.

Balanis, Constantine A. 1982. *Antenna Theory: Analysis and Design*. New York, NY: John Wiley.

Banzhaf, Wolfgang, Daida, Jason, Eiben, A. E., Garzon, Max H., Honavar, Vasant, Jakiela, Mark, and Smith, Robert E. (editors). 1999. *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13-17, 1999, Orlando, Florida USA*. San Francisco, CA: Morgan Kaufmann.

Banzhaf, Wolfgang, Nordin, Peter, Keller, Robert E., and Francone, Frank D. 1998. *Genetic Programming – An Introduction*. San Francisco, CA: Morgan Kaufmann and Heidelberg: dpunkt.

Banzhaf, Wolfgang, Poli, Riccardo, Schoenauer, Marc, and Fogarty, Terence C. 1998. *Genetic Programming: First European Workshop. EuroGP'98. Paris, France, April 1998 Proceedings. Paris, France. April l998*. Lecture Notes in Computer Science. Volume 1391. Berlin, Germany: Springer-Verlag.

Bennett, Forrest H III, Koza, John R., Shipman, James, and Stiffelman, Oscar. 1999. Building a parallel computer system for $18,000 that performs a half peta-flop per day. In Banzhaf, Wolfgang, Daida, Jason, Eiben, A. E., Garzon, Max H., Honavar, Vasant, Jakiela, Mark, and Smith, Robert E. (editors). 1999. *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13-17, 1999, Orlando, Florida USA*. San Francisco, CA: Morgan Kaufmann. Pages 1484 - 1490.

Boyd, S. P. and Barratt, C. H. 1991. *Linear Controller Design: Limits of Performance*. Englewood Cliffs, NJ: Prentice Hall.

Burke, Gerald J. 1992. *Numerical Electromagnetics Code — NEC-4: Method of Moments — User's Manual*. Lawrence Livermore National Laboratory report UCRL-MA-109338. Livermore, CA: Lawrence Livermore National Laboratory.

Callender, Albert and Stevenson, Allan Brown. 1939. *Automatic Control of Variable Physical Characteristics*. United States Patent 2,175,985. Filed February 17, 1936 in United States. Filed February 13, 1935 in Great Britain. Issued October 10, 1939 in United States.

Campbell, George A. 1917. *Electric Wave Filter*. Filed July 15, 1915. U. S. Patent 1,227,113. Issued May 22, 1917.

Comisky, William, Yu, Jessen, and Koza, John. 2000. Automatic synthesis of a wire antenna using genetic programming. *Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference, Las Vegas, Nevada*. Pages 179 - 186.

D'haeseleer, Patrik, Wen, Xiling, Fuhrman, Stefanie, and Somogyi, Roland. 1999. Linear modeling of mRNA expression levels during CNS development and injury. In Altman, Russ B. Dunker, A. Keith, Hunter, Lawrence, Klein, Teri E., and Lauderdale, Kevin (editors). *Pacific Symposium on Biocomputing '99*. Singapore: World Scientific. Pages 41 - 52.

Dorf, Richard C. and Bishop, Robert H. 1998. *Modern Control Systems*. Eighth edition. Menlo Park, CA: Addison-Wesley.

Holland, John H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Ann Arbor, MI: University of Michigan Press 1975. Second edition. Cambridge, MA: The MIT Press 1992.

Jones, Harry S. 1942. *Control Apparatus*. United States Patent 2,282,726. Filed October 25, 1939. Issued May 12, 1942.

Keane, Martin A., Yu, Jessen, and Koza, John R. 2000. Automatic synthesis of both the topology and tuning of a common parameterized controller for two families of plants using genetic programming. In Whitley, Darrell, Goldberg, David, Cantu-Paz, Erick, Spector, Lee, Parmee, Ian, and Beyer, Hans-Georg (editors). *GECCO-2000: Proceedings of the Genetic and Evolutionary Computation Conference, July 10 - 12, 2000, Las Vegas, Nevada*. San Francisco: Morgan Kaufmann Publishers. Pages 496 - 504. Kinnear, Kenneth E. Jr. (editor). 1994. *Advances in Genetic Programming*. Cambridge, MA: MIT Press.

Koza, John R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* Cambridge, MA: MIT Press.

Koza, John R. 1994a. *Genetic Programming II: Automatic Discovery of Reusable Programs.* Cambridge, MA: MIT Press.

Koza, John R. 1994b. *Genetic Programming II Videotape: The Next Generation*. Cambridge, MA: MIT Press.

Koza, John R., Banzhaf, Wolfgang, Chellapilla, Kumar, Deb, Kalyanmoy, Dorigo, Marco, Fogel, David B., Garzon, Max H., Goldberg, David E., Iba, Hitoshi, and Riolo, Rick. (editors). 1998. *Genetic Programming 1998: Proceedings of the Third Annual Conference*. San Francisco, CA: Morgan Kaufmann.

Koza, John R., Bennett III, Forrest H, Andre, David, and Keane, Martin A. 1999. *Genetic Programming III: Darwinian Invention and Problem Solving*. San Francisco, CA: Morgan Kaufmann.

Koza, John R., Bennett III, Forrest H, Andre, David, Keane, Martin A., and Brave Scott. 1999. *Genetic Programming III Videotape: Human-Competitive Machine Intelligence*. San Francisco, CA: Morgan Kaufmann.

Koza, John R., Deb, Kalyanmoy, Dorigo, Marco, Fogel, David B., Garzon, Max, Iba, Hitoshi, and Riolo, Rick L. (editors). 1997. *Genetic Programming 1997: Proceedings of the Second Annual Conference* San Francisco, CA: Morgan Kaufmann.

Koza, John R., Deb, Kalyanmoy, Dorigo, Marco, Fogel, David B., Garzon, Max, Iba, Hitoshi, and Riolo, Rick L. (editors). *Genetic Programming 1997: Proceedings of the Second Annual Conference, July 13–16, 1997, Stanford University*. San Francisco, CA: Morgan Kaufmann.

Koza, John R., Goldberg, David E., Fogel, David B., and Riolo, Rick L. (editors). 1996. *Genetic Programming 1996: Proceedings of the First Annual Conference, July 28-31, 1996, Stanford University*. Cambridge, MA: MIT Press.

Koza, John R., Keane, Martin A., Bennett, Forrest H III, Yu, Jessen, Mydlowec, William, and Stiffelman, Oscar. 1999. Automatic creation of both the topology and parameters for a robust controller by means of genetic programming. *Proceedings of the 1999 IEEE International Symposium on Intelligent Control, Intelligent Systems, and Semiotics*. Piscataway, NJ: IEEE. Pages 344 - 352.

Koza, John R., Keane, Martin A., Yu, Jessen, Bennett, Forrest H III, and Mydlowec, William. 2000. Automatic creation of human-competitive programs and controllers by means of genetic programming. *Genetic Programming and Evolvable Machines*. (1) 121 - 164.

Koza, John R., Keane, Martin A., Yu, Jessen, Bennett, Forrest H III, Mydlowec, William, and Stiffelman, Oscar. 1999. Automatic synthesis of both the topology and parameters for a robust controller for a non-minimal phase plant and a three-lag plant by means of genetic programming. *Proceedings of 1999 IEEE Conference on Decision and Control*. Pages 5292 - 5300.

Koza, John R., Keane, Martin A., Yu, Jessen, Mydlowec, William, and Bennett, Forrest H III. 2000a. Automatic synthesis of both the topology and parameters for a controller for a three-lag plant with a five-second delay using genetic programming. In Cagnoni, Stafano et al. (editors). *Real-World Applications of Evolutionary Computing. EvoWorkshops 2000. EvoIASP, Evo SCONDI, EvoTel, EvoSTIM, EvoRob, and EvoFlight, Edinburgh, Scotland, UK, April 2000, Proceedings*. Lecture Notes in Computer Science. Volume 1803. Berlin, Germany: Springer-Verlag. Pages 168 - 177. ISBN 3-540-67353-9.

Koza, John R., Keane, Martin A., Yu, Jessen, Mydlowec, William, and Bennett, Forrest H III. 2000b. Automatic synthesis of both the control law and parameters for a controller for a three-lag plant with five-second delay using genetic programming and simulation techniques. In *Proceedings of the 2000 American Control Conference, Chicago, Illinois, June 28 - 30, 2000*. Evanston, IL: American Automatic Control Council. Pages 453-459.

Koza, John R., Mydlowec, William, Lanza, Guido, Yu, Jessen, and Keane, Martin A. 2000. *Reverse Engineering and Automatic Synthesis of Metabolic Pathways from Observed Data Using Genetic Programming*. Stanford Medical Informatics Technical Report SMI-2000-0851.

Koza, John R., and Rice, James P. 1992. *Genetic Programming: The Movie*. Cambridge, MA: MIT Press.

Koza, John R., Yu, Jessen, Keane, Martin A., and Mydlowec, William. 2000. Evolution of a controller with a free variable using genetic programming. In Poli, Riccardo, Banzhaf, Wolfgang, Langdon, William B., Miller, Julian, Nordin, Peter, and Fogarty, Terence C. 2000. *Genetic Programming: European Conference, EuroGP 2000, Edinburgh, Scotland, UK, April 2000, Proceedings*. Lecture Notes in Computer Science. Volume 1802. Berlin, Germany: Springer-Verlag. Pages 91 - 105. ISBN 3-540-67339-3.

Laing, Shoudan, Fuhrman, Stefanie, and Somogyi, Roland. 1998. REVEAL: A general reverse engineering algorithm for inference of genetic network architecture. In Altman, Russ B. Dunker, A. Keith, Hunter, Lawrence, and Klein, Teri E. (editors). *Pacific Symposium on Biocomputing '98*. Singapore: World Scientific. Pages 18 - 29.

Langdon, William B. 1998. *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!* Amsterdam: Kluwer.

Linden, Derek S. 1997. *Automated Design and Optimization of Wire Antennas Using Genetic Algorithms*. Ph.D. thesis. Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.

Loomis, William F. and Sternberg, Paul W. 1995. Genetic networks. *Science*. Pages 269. 649. August 4, 1995.

McAdams, Harley H. and Shapiro, Lucy. 1995. Circuit simulation of genetic networks. *Science*. 269. Pages 650-656. August 4, 1995.

Mittenthal, Jay E., Ao Yuan, Bertrand Clarke, and Scheeline, Alexander. 1998. Designing metabolism: Alternative connectivities for the pentose phosphate pathway. *Bulletin of Mathematical Biology*. 60: 815 - 856.

Mydlowec, William and Koza, John. 2000. Use of time-domain simulations in automatic synthesis of computational circuits using genetic programming. *Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference, Las Vegas, Nevada*. Pages 187 - 197.

Poli, Riccardo, Nordin, Peter, Langdon, William B., and Fogarty, Terence C. 1999. *Genetic Programming: Second European Workshop. EuroGP'99. Proceedings*. Lecture Notes in Computer Science. Volume 1598. Berlin, Germany: Springer-Verlag.

Quarles, Thomas, Newton, A. R., Pederson, D. O., and Sangiovanni-Vincentelli, A. 1994. *SPICE 3 Version 3F5 User's Manual*. Department of Electrical Engineering and Computer Science, University of California. Berkeley, CA. March 1994.

Sterling, Thomas L., Salmon, John, and Becker, Donald J., and Savarese, Daniel F. 1999. *How to Build a Beowulf: A Guide to Implementation and Application of PC Clusters*. Cambridge, MA: MIT Press.

Stutzman, Warren. L. and Thiele, Gary A. 1998. *Antenna Theory and Design*. Second edition. New York, NY: John Wiley.

Tomita, Masaru, Hashimoto, Kenta, Takahashi, Kouichi, Shimizu, Thomas Simon, Matsuzaki, Yuri, Miyoshi, Fumihiko, Saito, Kanako, Tanida, Sakura, Yugi, Katsuyuki, Venter, J. Craig, Hutchison, Clyde A. III. 1999. E-CELL: Software environment for whole cell simulation. *Bioinformatics*. Volume 15 (1) 72-84.

Uda, S. 1926. Wireless beam of short electric waves. *Journal of the IEE (Japan)*. March 1926. 273 - 282.

Uda, S. 1927. Wireless beam of short electric waves. *Journal of the IEE (Japan)*. March 1927.1209-1219.

Voit, Eberhard O. 2000. *Computational Analysis of Biochemical Systems*. Cambridge: Cambridge University Press.

Webb, Edwin C. 1992. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. San Diego, CA: Academic Press.

Whitley, Darrell, Goldberg, David, Cantu-Paz, Erick, Spector, Lee, Parmee, Ian, and Beyer, Hans-Georg (editors). 2000. *GECCO-2000: Proceedings of the Genetic and Evolutionary Computation Conference, July 10 - 12, 2000, Las Vegas, Nevada*. San Francisco: Morgan Kaufmann Publishers.

Wong, Man Leung and Leung, Kwong Sak. 2000. *Data Mining Using Grammar Based Genetic Programming and Applications*. Amsterdam: Kluwer Academic Publishers. ISBN: 0-7923-7746-X

Yagi, H. 1928. Beam transmission of ultra short waves. *Proceedings of the IRE*. 26: 714-741. June 1928.

Yu, Jessen, Keane, Martin A., and Koza, John R. 2000. Automatic design of both topology and tuning of a common parameterized controller for two families of plants using genetic programming. In *Proceedings of Eleventh IEEE International Symposium on Computer-Aided Control System Design (CACSD) Conference and*

Ninth IEEE International Conference on Control Applications (CCA) Conference, Anchorage, Alaska, September 25 - 27, 2000. In Press.

Yuh, Chiou-Hwa, Bolouri, Hamid, and Davidson, Eric H. 1998. Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science*. 279. Pages 1896 - 1902.

Zobel, Otto Julius. 1925. *Wave Filter*. Filed January 15, 1921. U. S. Patent 1,538,964. Issued May 26, 1925.