

METABOLIC PATHWAYS (NETWORKS OF CHEMICAL REACTIONS)

GENETIC NETWORKS

THE *TOPOLOGY* OF A NETWORK OF CHEMICAL REACTIONS

- the total number of reactions in the network,
- the number of substrate(s) consumed by each reaction,
- the number of product(s) produced by each reaction,
- the pathways supplying the substrate(s) (either from external sources or other reactions in the network) to each reaction,
- the pathways dispersing each reaction's product(s) (either to other reactions or external outputs), and
- an indication of which enzyme (if any) acts as a catalyst for a particular reaction

THE *SIZING* FOR A NETWORK OF CHEMICAL REACTIONS

- all the numerical values associated with the network (e.g., the rates of each reaction)

OUR APPROACH

- **establishing a representation for chemical networks involving symbolic expressions (S-expressions) and program trees that can be progressively bred (and improved) by means of genetic programming,**
- **converting each individual program tree in the population into an analog electrical circuit representing the network of chemical reactions,**
- **obtaining the behavior of the individual network of chemical reactions by simulating the corresponding electrical circuit,**
- **defining a fitness measure that measures how well the behavior of an individual network matches the observed time-domain data concerning concentrations of final product substance(s), and**
- **using the fitness measure to enable genetic programming to breed an improved population of program trees.**

FIVE DIFFERENT REPRESENTATIONS

- **Reaction Network:** The blocks represent chemical reactions and the directed lines represent flows of substances between reactions.
- **Program Tree:** A network of chemical reactions can also be represented as a program tree whose internal points are functions and external points are terminals. This representation enables genetic programming to breed a population of programs in a search for a network of chemical reactions whose time-domain behavior concerning concentrations of final product substance(s) closely matches observed data.
- **Symbolic Expression:** A network of chemical reactions can also be represented as a symbolic expression (S-expression) in the style of the LISP programming language. This representation is used internally by the run of genetic programming.

FIVE DIFFERENT REPRESENTATIONS — CONTINUED

- **System of Non-Linear Differential Equations: A network of chemical reactions can also be represented as a system of non-linear differential equations.**
- **Analog Electrical Circuit: A network of chemical reactions can also be represented as an analog electrical circuit. Representation of a network of chemical reactions as a circuit facilitates simulation of the network's time-domain behavior.**

ILLUSTRATIVE PROBLEM NO. 1 — PHOSPHOLIPID CYCLE

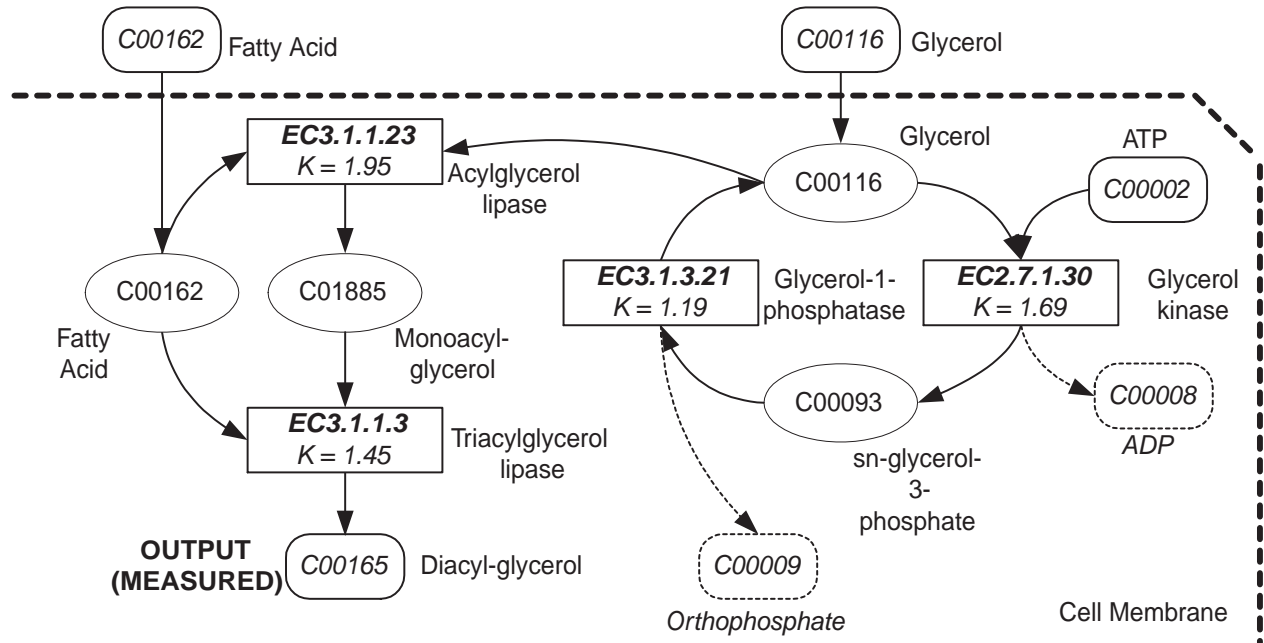
- **4 reactions that are part of the phospholipid cycle, as presented in the E-CELL cell simulation model**
- **External inputs**
 - glycerol (C00116)
 - fatty acid (C00162).
 - cofactor ATP(C00002)
- **Network's final product**
 - diacyl-glycerol (C00165)
- **Catalysts**
 - Glycerol kinase (EC2.7.1.30),
 - Glycerol-1-phosphatase (EC3.1.3.21),
 - Acylglycerol lipase (EC3.1.1.23), and
 - Triacylglycerol lipase (EC3.1.1.3)
- **2 intermediate substances**
 - sn-Glycerol-3-Phosphate (C00093)
 - Monoacyl-glycerol (C01885)

ILLUSTRATIVE PROBLEM NO. 1 — PHOSPHOLIPID CYCLE

INTERESTING TOPOLOGY

- **2 instances of a bifurcation point (where one substance is distributed to two different reactions)**
 - External supply of fatty acid (C00162) is distributed
 - External supply of glycerol (C00116) is distributed
- **1 instance of an accumulation point (where one substance is accumulated from two sources)**
 - glycerol (C00116) is externally supplied and
 - glycerol (C00116) is produced by the reaction catalyzed by Glycerol-1-phosphatase (EC3.1.3.21)
- **1 internal feedback loop (in which a substance is both consumed and produced)**
 - Glycerol (C00116) is consumed (in part) by the reaction catalyzed by Glycerol kinase (EC2.7.1.30).
 - This reaction, in turn, produces an intermediate substance, sn-Glycerol-3-Phosphate (C00093).
 - This intermediate substance is, in turn, consumed by the reaction catalyzed by Glycerol-1-phosphatase (EC3.1.3.21).
 - That reaction, in turn, produces glycerol (C00116).

FOUR REACTIONS FROM THE PHOSPHOLIPID CYCLE



ILLUSTRATIVE PROBLEM NO. 2 — SYNTHESIS AND DEGRADATION OF KETONE BODIES

- **3 reactions**
- **External inputs**
 - Acetoacetyl-CoA
 - Acetyl-CoA
- **Final product**
 - Acetoacetate.
- **Catalysts**
 - 3-oxoacid CoA-transferase (EC 2.8.3.5)
 - Hydroxymethylglutaryl-CoA synthase (EC 4.1.3.5)
 - Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)
- **1 intermediate substance**
 - INT-1

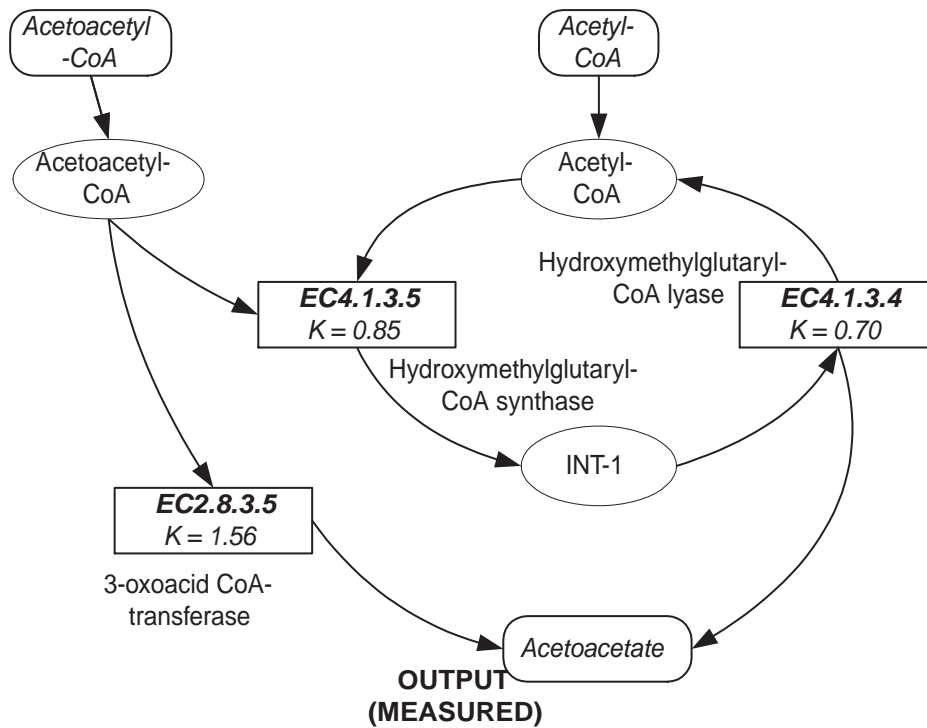
ILLUSTRATIVE PROBLEM NO. 2 — SYNTHESIS AND DEGRADATION OF KETONE BODIES

3 NOTEWORTHY TOPOLOGICAL FEATURES

- **1 instance of a bifurcation point (where one substance is distributed to two different reactions)**
 - Acetoacetyl-CoA
- **2 accumulation points**
 - Acetyl-CoA is an externally supplied substance and is produced by the reaction catalyzed by Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)
 - Acetoacetate is produced by the reaction catalyzed by 3-oxoacid CoA-transferase (EC 2.8.3.5) and by the reaction catalyzed by Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)
- **1 internal feedback loop (in which a substance is both consumed and produced) • Acetyl-CoA is consumed by the reaction catalyzed by Hydroxymethylglutaryl-CoA synthase (EC 4.1.3.5).**

- This reaction, in turn, produces an intermediate substance (INT-1)
- This intermediate substance is, in turn, consumed by the reaction catalyzed by Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4).
- That reaction, in turn, produces Acetyl-CoA.

THREE REACTIONS INVOLVED IN THE SYNTHESIS AND DEGRADATION OF KETONE BODIES



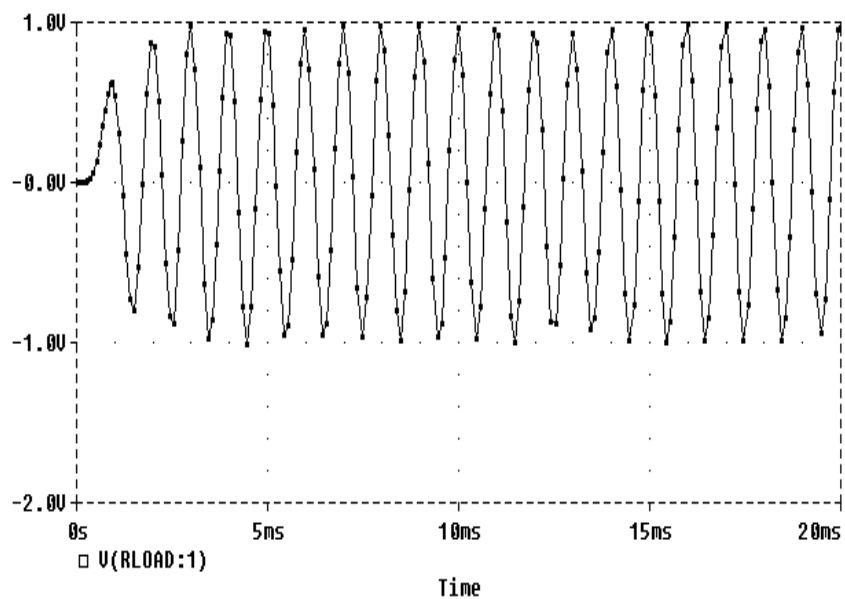
ARCHITECTURE-ALTERING OPERATIONS IN GP

- **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**
 - the total number of automatically defined functions,
 - the number of arguments (if any) possessed by each automatically defined function, and
 - 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.

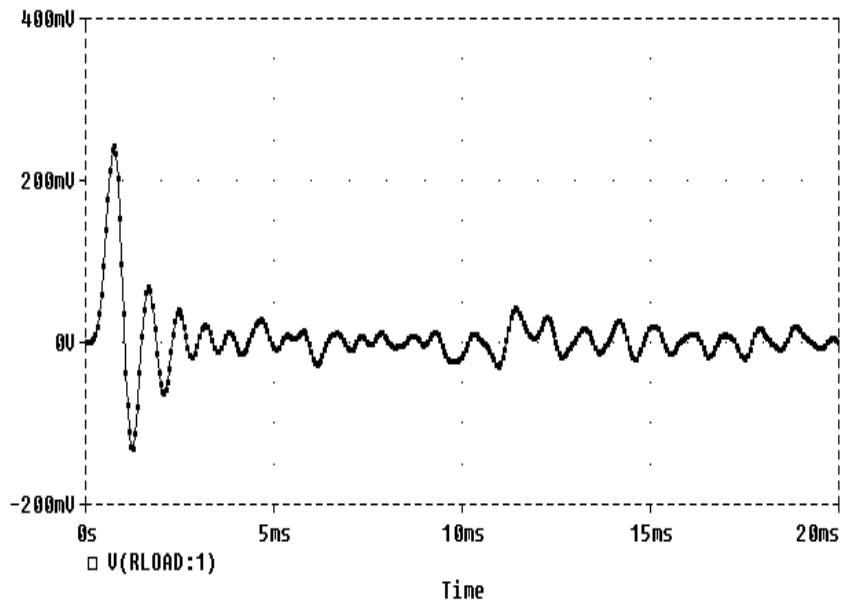
AUTOMATIC SYNTHESIS OF ANALOG ELECTRICAL CIRCUITS

LOWPASS FILTER CIRCUIT

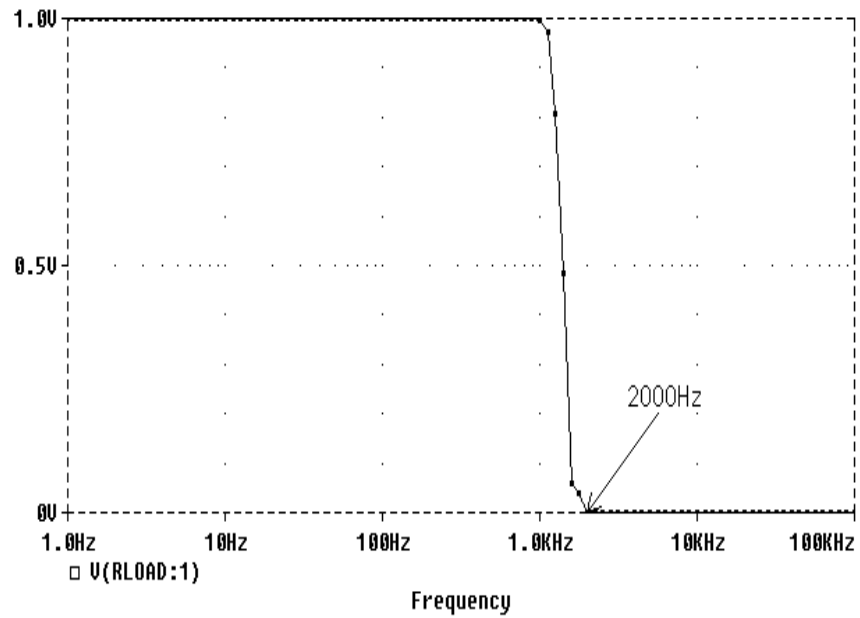
TIME DOMAIN BEHAVIOR OF A LOWPASS FILTER TO A 1,000 HZ SINUSOIDAL INPUT SIGNAL



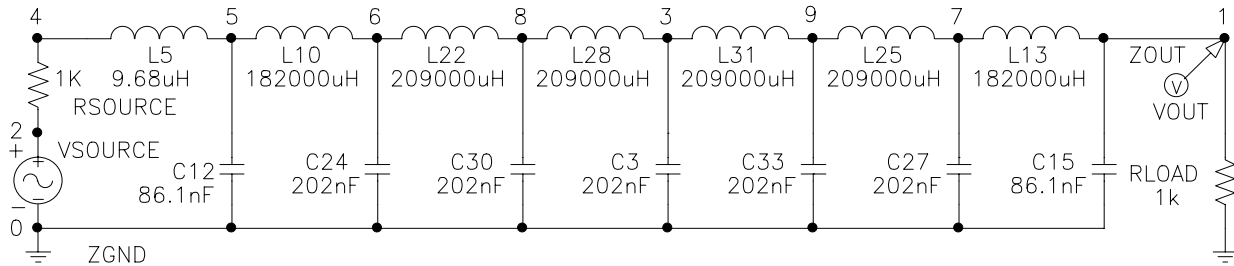
TIME DOMAIN BEHAVIOR OF A LOWPASS FILTER TO A 2,000 HZ SINUSOIDAL INPUT SIGNAL



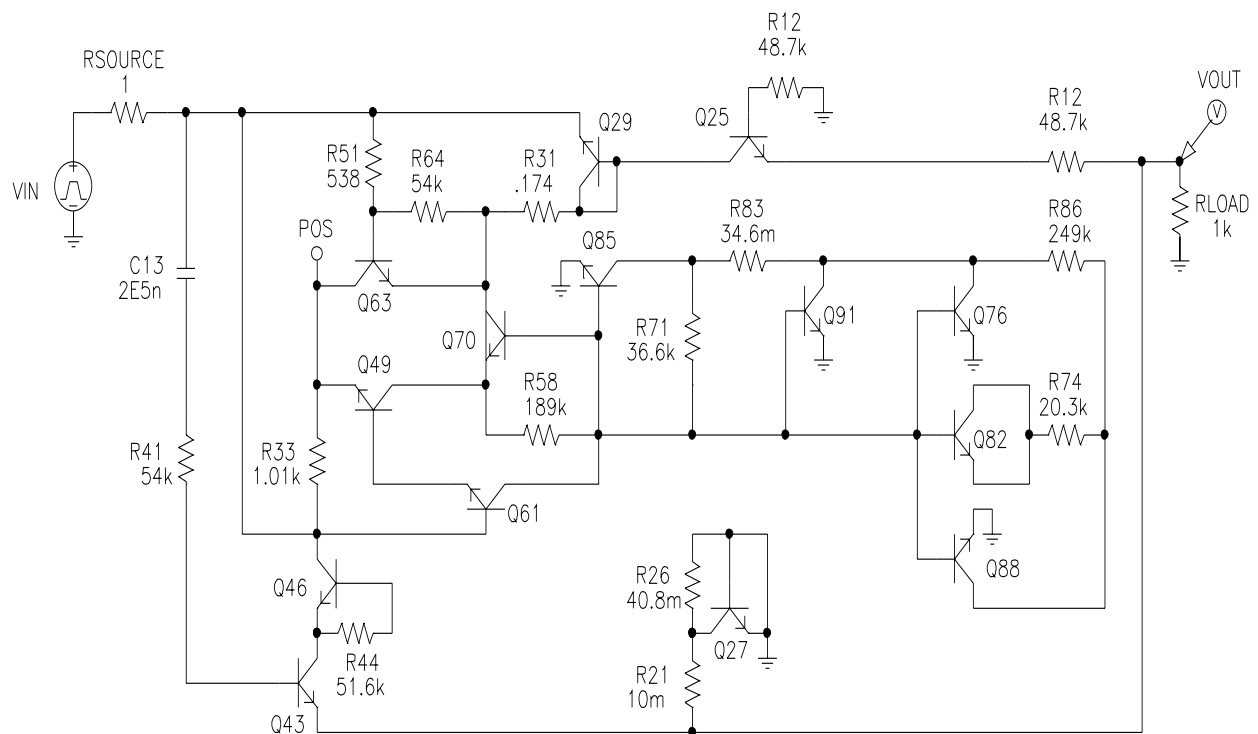
FREQUENCY DOMAIN BEHAVIOR OF A LOWPASS FILTER



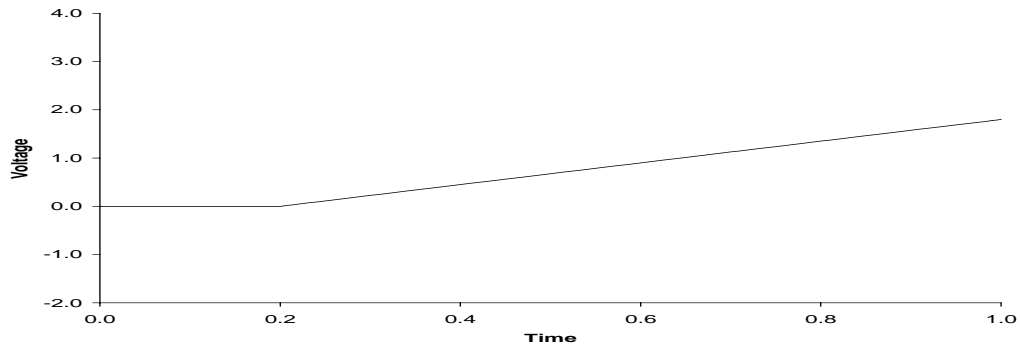
LOWPASS FILTER CREATED BY GENETIC PROGRAMMING THAT INFRINGES ON GEORGE CAMPBELL'S PATENT



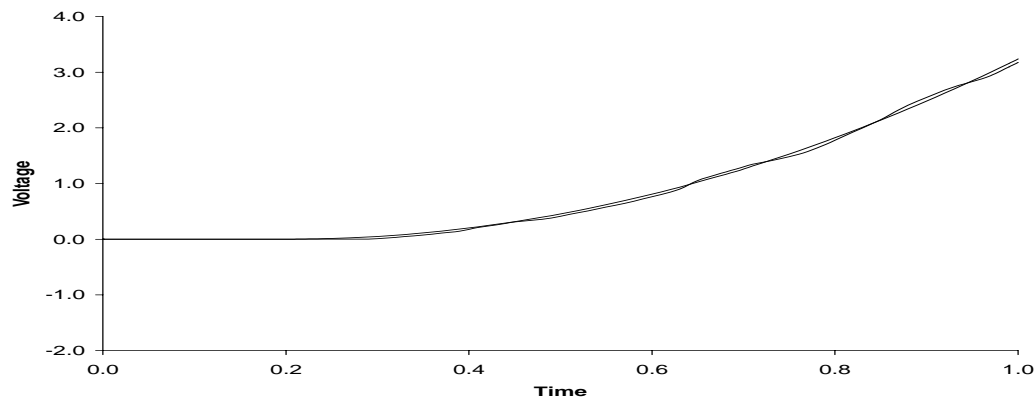
SQUARING COMPUTATIONAL CIRCUIT CREATED BY GENETIC PROGRAMMING



RISING RAMP — 1 OF 4 TIME-DOMAIN SIGNALS USED TO CREATE SQUARING COMPUTATIONAL CIRCUIT

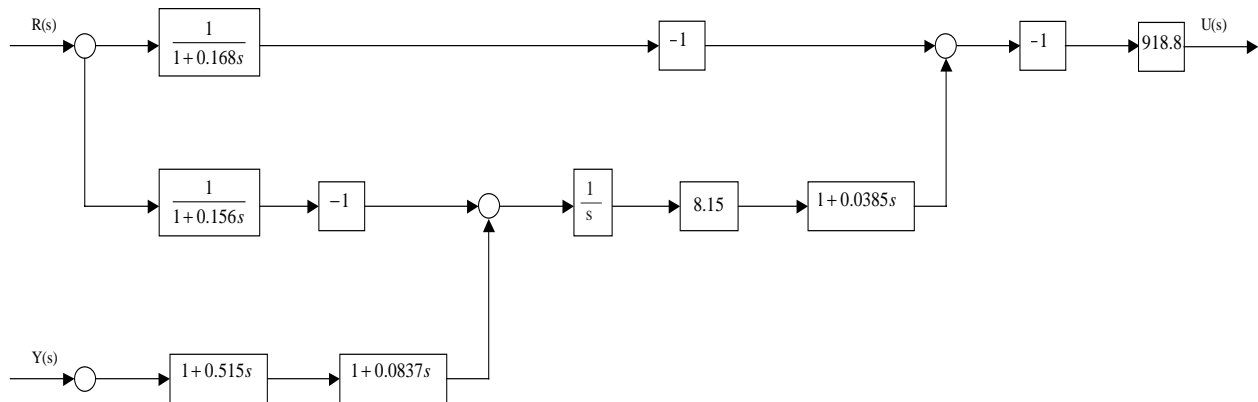


OUTPUT FOR RISING RAMP INPUT FOR SQUARING CIRCUIT



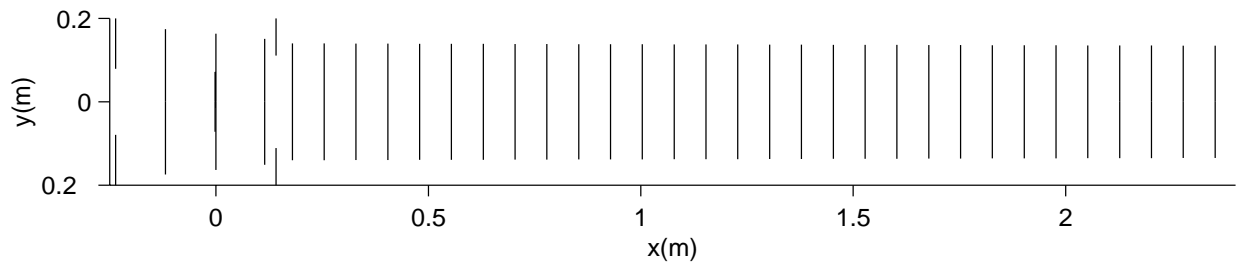
AUTOMATIC SYNTHESIS OF CONTROLLERS

EVOLVED CONTROLLER THAT INFRINGES ON JONES' PATENT



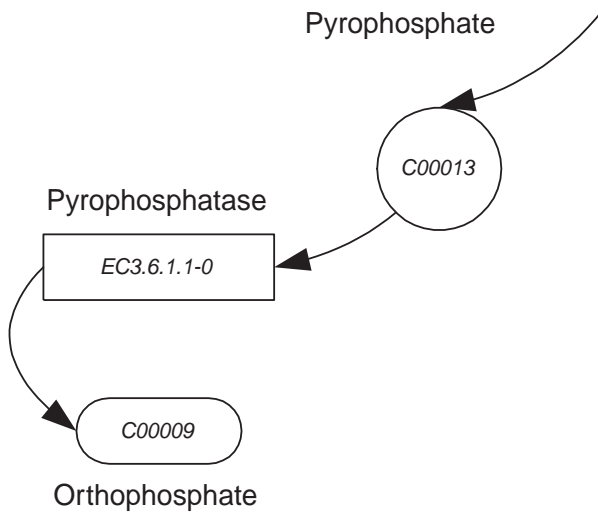
AUTOMATIC SYNTHESIS OF ANTENNAS

ANTENNA DESIGN CREATED BY GENETIC PROGRAMMING



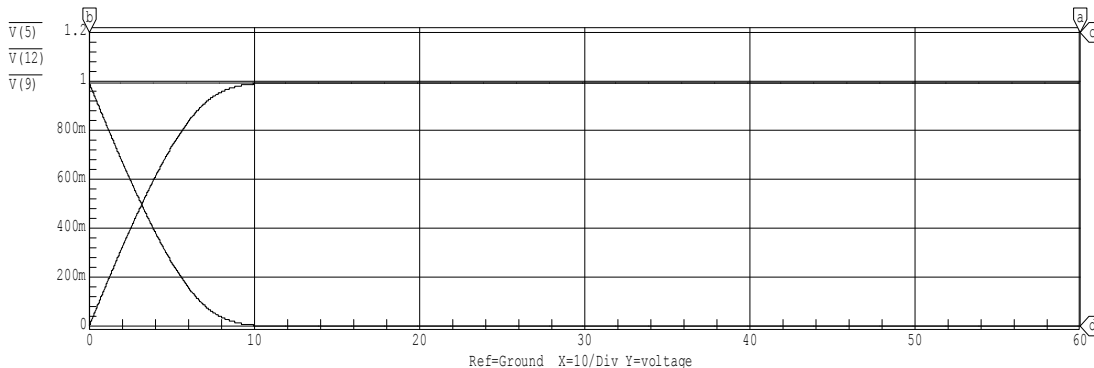
ONE-SUBSTRATE, ONE-PRODUCT CHEMICAL REACTION

- One chemical (the *substrate*) is transformed into another chemical (the *product*) under control of a catalyst



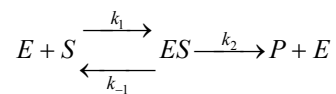
CHANGING CONCENTRATIONS OF SUBSTANCES IN AN ILLUSTRATIVE ONE-SUBSTRATE, ONE-PRODUCT REACTION

Xa: 60.00 Xb: 0.000 a-b: 60.00 freq: 16.67m
Yc: 1.200 Yd: 0.000 c-d: 1.200



CHEMICAL REACTIONS

- **The action of an enzyme (catalyst) in a one-substrate chemical reaction can be viewed as two-step process in which the enzyme E first binds with the substrate S at a rate k_1 to form ES. The formation of the product P from ES then occurs at a rate k_2 . The reverse reaction (for the binding of E with S) in which ES dissociates into E and S, occurs at a rate of k_{-1}**



CHEMICAL REACTIONS

• The concentrations of substrates, products, intermediate substances, and catalysts participating in reactions are modeled by various rate laws, including

- first-order rate laws,
- second-order rate laws, power laws, and
- Michaelis-Menten equations

• Michaelis-Menten rate law for a one-substrate chemical reaction is

$$\frac{d[P]}{dt} = \frac{k_2[E]_0[S]_t}{[S]_t + K_m}.$$

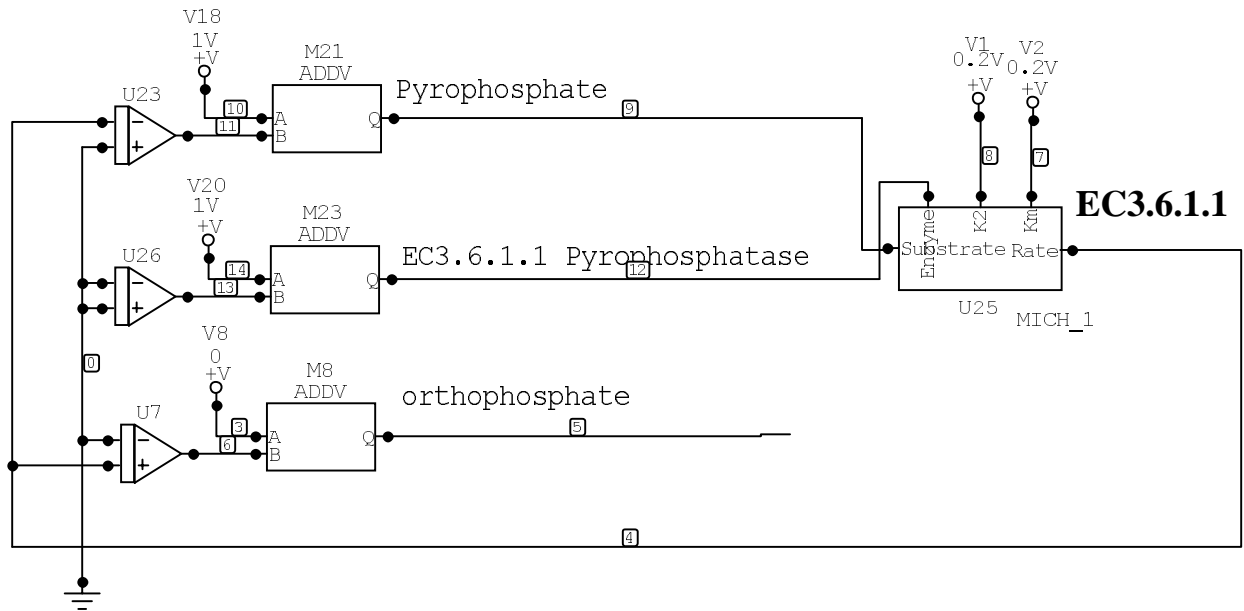
$$K_m = \frac{k_{-1} + k_2}{k_1}.$$

• Psuedo-first-order rate law

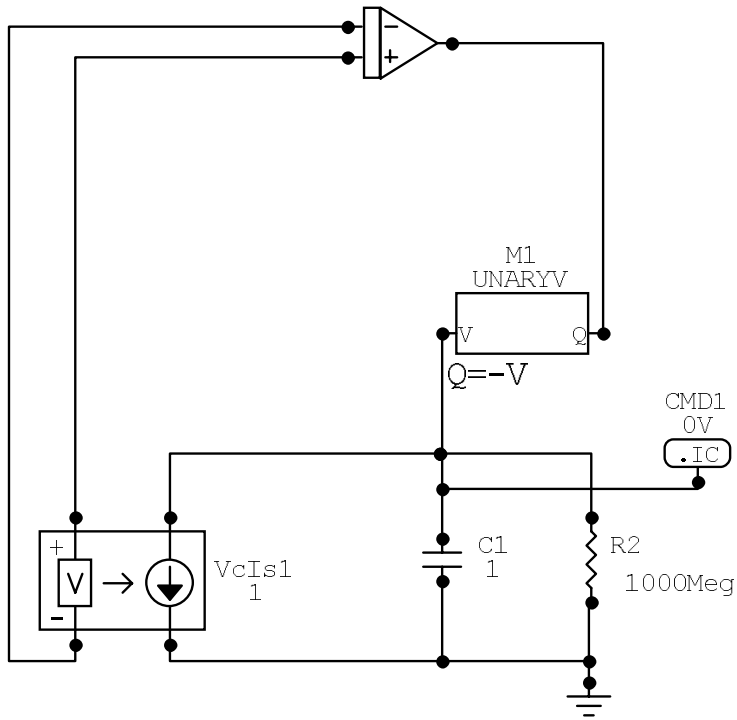
$$\frac{d[P]}{dt} = \frac{k_2[E]_0[S]_t}{K_m} = k_{new}[E]_0[S]_t,$$

$$k_{new} = \frac{k_2}{K_m}.$$

ELECTRICAL CIRCUIT REPRESENTING THE ILLUSTRATIVE ONE-SUBSTRATE-ONE-PRODUCT ENZYMATIC REACTION

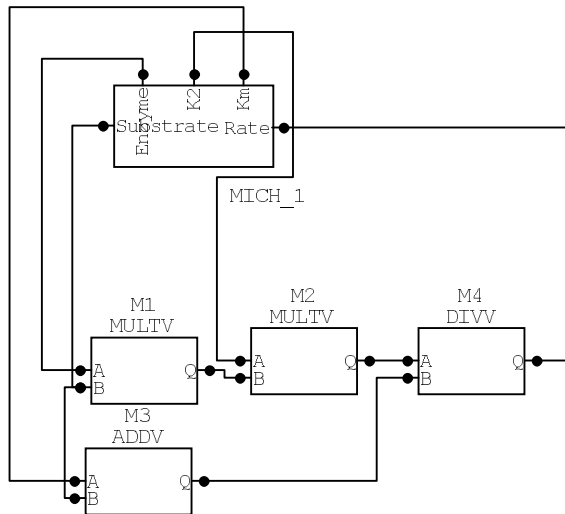


SUM-INTEGRATOR



SUBCIRCUIT FOR ONE-SUBSTRATE MICHAELIS-MENTEN EQUATION

MICH_1



Subcircuit definition in SPICE for the one-substrate Michaelis-Menten equation

MICH_1

***NETLIST FOR MICHAELIS-MENTEN**

MICH_1

XXM4 4 3 2 XDIVV

XXM3 6 5 3 XADDV

XXM2 7 8 4 XMULTV

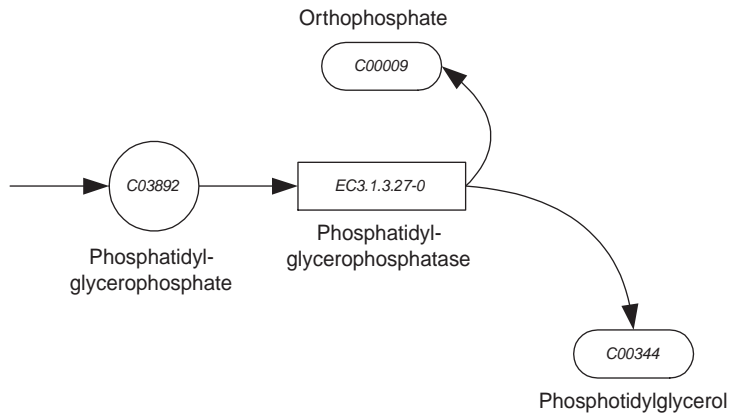
XXM1 9 5 8 XMULTV

.SAVE V(2) V(3) V(4) V(5) V(6)

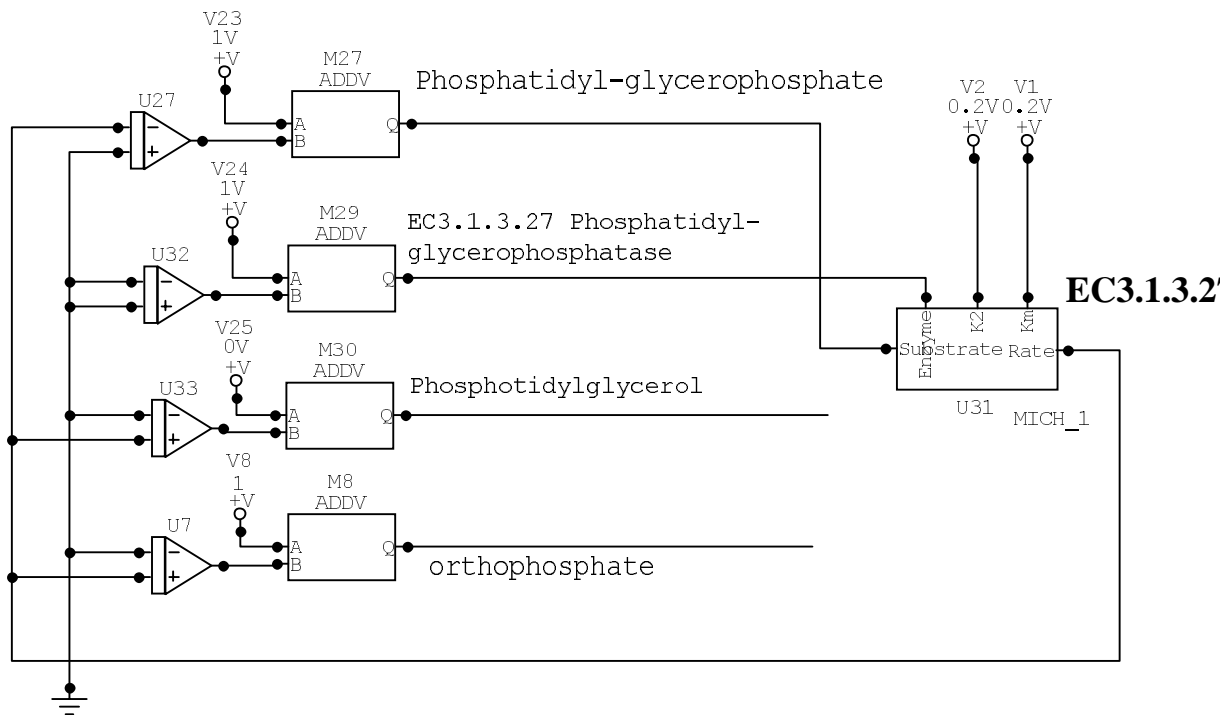
V(7) V(8) V(9)

.END

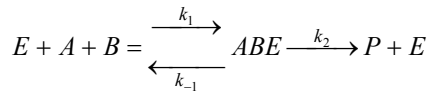
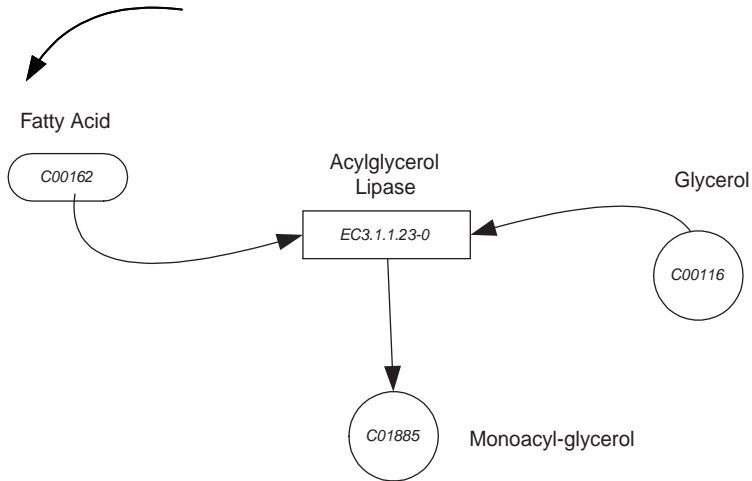
ONE-SUBSTRATE, TWO-PRODUCT REACTION



CIRCUIT FOR ILLUSTRATIVE ONE-SUBSTRATE, TWO-PRODUCT CHEMICAL REACTION



TWO-SUBSTRATE, ONE-PRODUCT REACTION



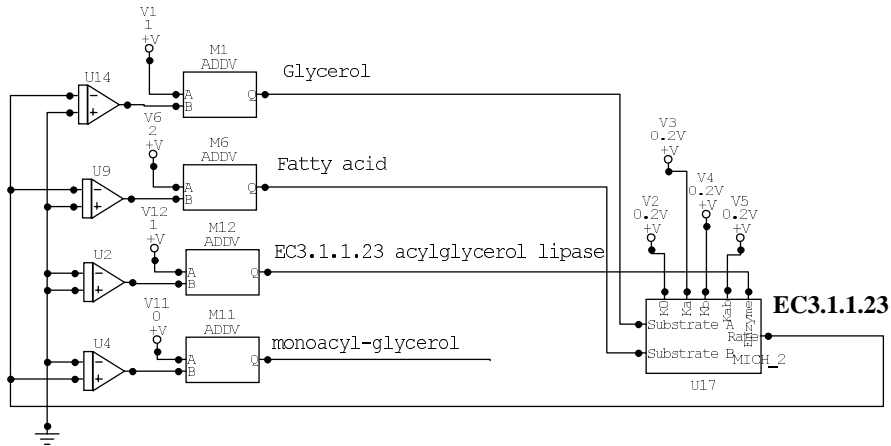
- **Michaelis-Menten rate law for a two-substrate chemical reaction is**

$$Rate_t = \frac{[E]_0}{\frac{1}{K_0} + \frac{1}{K_A[A]_t} + \frac{1}{K_B[B]_t} + \frac{1}{K_{AB}[A]_t[B]_t}}$$

- **When $k_{-1} \sim 0$ and $k_{-1} \ll k_1 \ll k_2$, it is often satisfactory to use a psuedo-second-order rate law such as**

$$Rate_t = k_1[A][B][E]$$

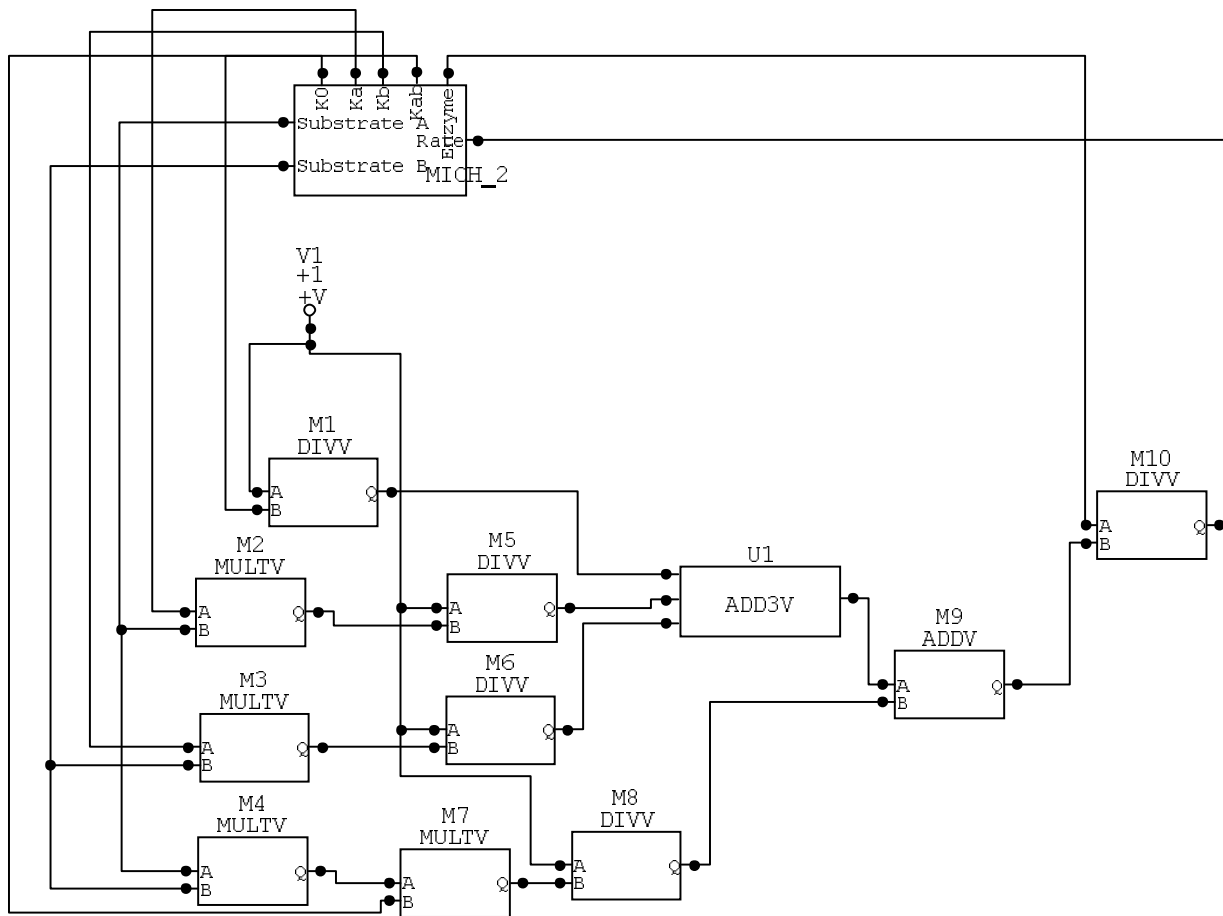
CIRCUIT FOR TWO-SUBSTRATE, ONE-PRODUCT CHEMICAL REACTION



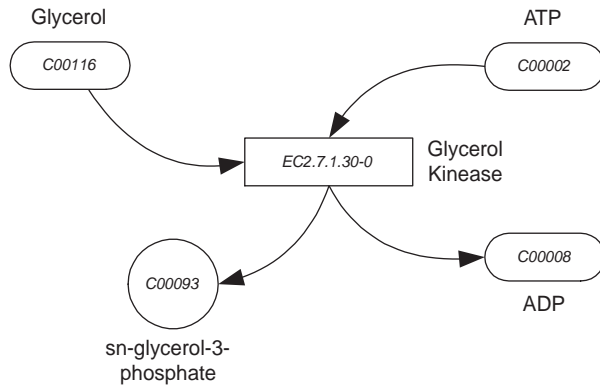
TWO-SUBSTRATE MICHAELIS-MENTEN EQUATION MICH_2

$$Rate_t = \frac{[E]_0}{\frac{1}{K_0} + \frac{1}{K_A[A]_t} + \frac{1}{K_B[B]_t} + \frac{1}{K_{AB}[A]_t[B]_t}}$$

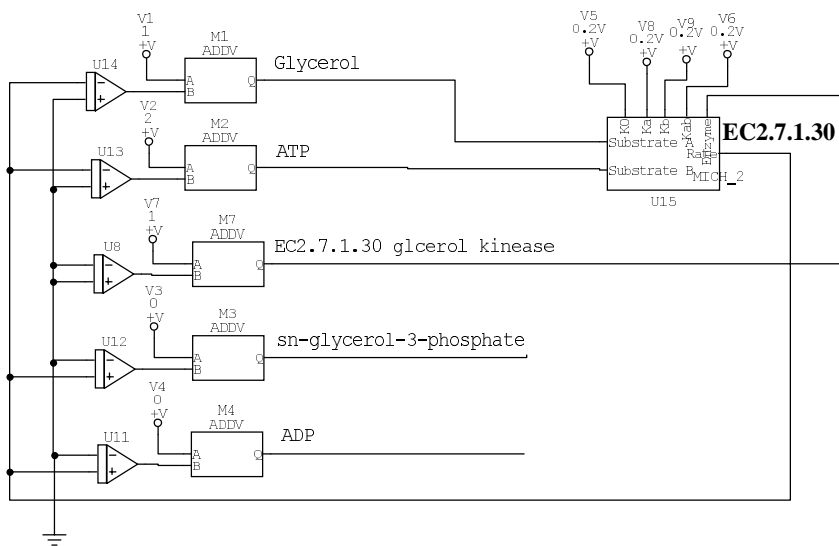
SUBCIRCUIT FOR TWO-SUBSTRATE MICHAELIS-MENTEN EQUATION MICH_2



TWO-SUBSTRATE, TWO-PRODUCT REACTION



ELECTRICAL CIRCUIT REPRESENTING A TWO-SUBSTRATE, TWO-PRODUCT ENZYMIC REACTION



REPertoire OF FUNCTIONS IN PROGRAM TREE

FOUR CHEMICAL REACTION FUNCTIONS

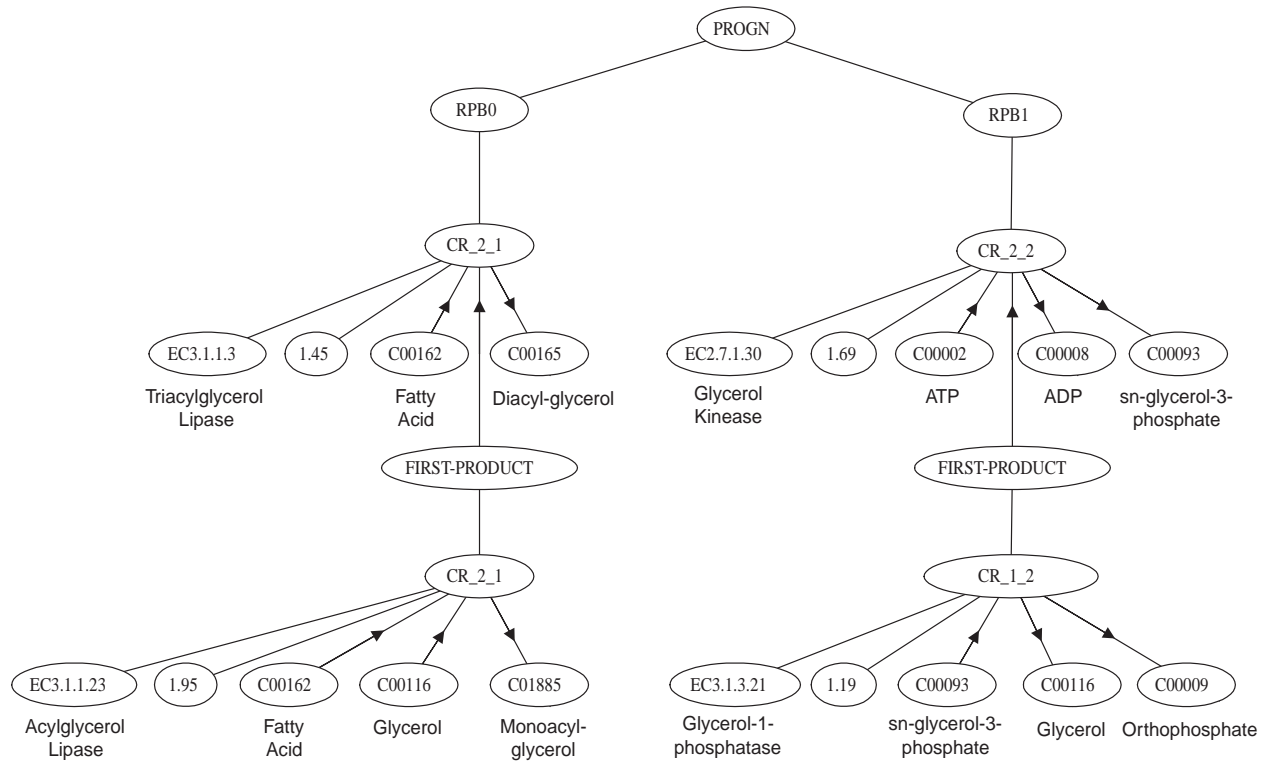
Function	Substr ates	Product s	Arit y
CR_1_1	1	1	4
CR_1_2	1	2	5
CR_2_1	2	1	5
CR_2_2	2	2	6

- Each chemical reaction function returns a list (of length 1 or 2) composed of the reaction's one or two products.
- The one-argument **FIRST-PRODUCT** function returns the first of the one or two products produced by the chemical reaction function designated by its argument.
- The one-argument **SECOND-PRODUCT** function returns the second of the two products (or, the first product, if the reaction produces only one product).

REPertoire OF TERMINALS IN THE PROGRAM TREE

- **Substances**
 - **externally supplied input substances**
 - **intermediate substances created by reactions**
 - **output substances**
- **Enzymes**
- **Numerical constants for the rate of the reactions**

PROGRAM TREE CORRESPONDING TO METABOLIC PATHWAY FOR PHOSPHOLIPID CYCLE



REPRESENTATION OF PHOSPHOLIPID CYCLE AS A SYMBOLIC EXPRESSION

```
(PROGn
  (CR_2_1
    EC3_1_1_3
    1.45
    C00162
    (FIRST-PRODUCT
      (CR_2_1
        EC3_1_1_23
        1.95
        C00162
        C00116
        C01885
      )
    )
    C00165
  )
  (CR_2_2
    EC2_7_1_30
    1.69
    C00002
    (FIRST-PRODUCT
      (CR_1_2
        EC3_1_3_21
        C00093
        C00116
        C00009
      )
    )
    C00008
    C00093
  )
)
```

REPRESENTATION OF PHOSPHOLIPID CYCLE AS A SYSTEM OF NON-LINEAR DIFFERENTIAL EQUATIONS

- **Supply of the network's final product, diacyl-glycerol (C00165)**

$$\frac{d[C00165]}{dt} = 1.45[C00162][C01885][EC 3.1.1.3]$$

- **Supply and consumption of the intermediate substance Monoacyl-glycerol (C01885)**

$$\frac{d[C01885]}{dt} = 1.95[C00162][C00116][EC 3.1.1.23] - 1.45[C00162][C01885][EC 3.1.1.3]$$

- **Supply and consumption of the intermediate substance sn-Glycerol-3-Phosphate (C00093) in the internal feedback loop**

$$\frac{d[C00093]}{dt} = 1.69[C00116][C00002][EC 2.7.1.30] - 1.19[C00093][EC 3.1.3.21]$$

- **Supply and consumption of cofactor ATP (C00002)**

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

REPRESENTATION OF PHOSPHOLIPID CYCLE AS A SYSTEM OF NON-LINEAR DIFFERENTIAL EQUATIONS — CONTINUED

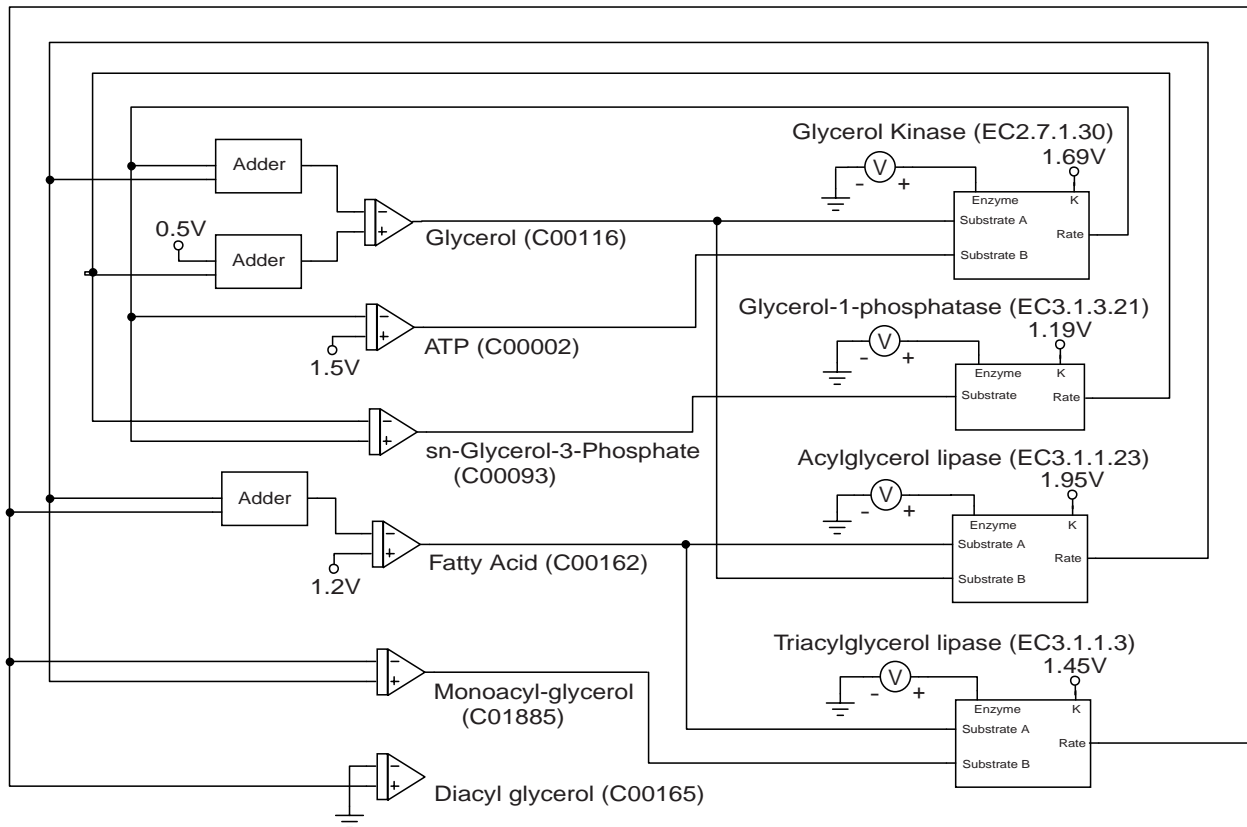
- **Supply and consumption of fatty acid (C00162)**

$$\frac{d[C00162]}{dt} = 1.2 - 1.95[C00162][C00116][EC3.1.1.23] - 1.45[C00162][C01885][EC3.1.1.3]$$

- **Supply, consumption, and production of glycerol (C00116)**

$$\frac{d[C00116]}{dt} = 0.5 + 1.19[C00093][EC3.1.3.21] - 1.69[C00116][C00002][EC2.7.1.30] - 1.95[C00162][C00116][EC3.1.1.23]$$

ELECTRICAL CIRCUIT CORRESPONDING TO THE METABOLIC PATHWAY FOR PHOSPHOLIPID CYCLE



PREPARATORY STEPS FOR GENETIC PROGRAMMING

PROGRAM ARCHITECTURE

- **Each program tree in the initial random population (generation 0) has one result-producing branch.**
- **In subsequent generations, the architecture-altering operations (patterned after gene duplication and gene deletion in nature) may add and delete result-producing branches to particular individual program trees in the population.**
- **Each program tree may have four result-producing branches.**

FUNCTION SET

$F = \{\text{CR_1_1}, \text{CR_1_2}, \text{CR_2_1}, \text{CR_2_2}, \text{FIRST-PRODUCT}, \text{SECOND-PRODUCT}\}.$

TERMINAL SET

$T = \{\mathcal{R}, \text{INT_1}, \text{INT_2}, \text{INT_3}, \text{C00116}, \text{C00162}, \text{C00002}, \text{C00165}, \text{EC2_7_1_30}, \text{EC3_1_3_21}, \text{EC3_1_1_23}, \}$

FITNESS MEASURE

- Concentrations of each of the four enzymes (EC2.7.1.30, EC3.1.3.21, EC3.1.1.23, and EC3.1.1.3) are varied in accordance with 6 different time series patterns over 30 half-second time steps:
- Each individual chemical reaction network is exposed to 9 test cases. Thus, there are 270 fitness cases (9 test cases, each consisting of 30 time steps).
- Each of the nine test cases is constructed by choosing four different time series from the above set of six time series as the concentration for the four enzymes (EC2.7.1.30, EC3.1.3.21, EC3.1.1.23, and EC3.1.1.3)

VARIATION IN THE LEVELS OF THE FOUR ENZYMES FOR THE NINE TEST CASES.

Test case	EC2.7.1.30	EC3.1.3.21	EC3.1.1.23	EC3.1.1.3
1	Slope-Up	Sawtooth	Step-Down	Step-Up
2	Slope-Down	Step-Up	Sawtooth	Step-Down
3	Step-Down	Slope-Up	Slope-Down	Step-Up
4	Step-Up	Slope-Down	Step-Up	Step-Down
5	Sawtooth	Step-Down	Slope-Up	Step-Up
6	Sawtooth	Step-Down	Knock-Out	Slope-Up
7	Sawtooth	Knock-Out	Slope-Up	Step-Down
8	Knock-Out	Step-Down	Slope-Up	Sawtooth
9	Step-Down	Slope-Up	Sawtooth	Knock-Out

FITNESS MEASURE — CONTINUED

- **There is a total of 270 data points. The data was obtained from the E-CELL cell simulation model**
- **The concentrations of all intermediate substances and the network's final product are 0 at time step 0.**
- **Glycerol (C00116), Fatty acid (C00162), and ATP (C00002) are externally supplied at a constant rate that is not subject to evolutionary change during the run**

FITNESS MEASURE — CONTINUED

RATES FOR THREE EXTERNALLY SUPPLIED SUBSTANCES

Substance	Rate
Glycerol (C00116)	0.5
Fatty acid (C00162)	1.2
ATP (C00002)	1.5

- Fitness is the sum, over the 270 fitness cases, of the absolute value of the difference between the concentration of the end product of the individual reaction network (diacylglycerol C00165 for the first network and Acetoacetate for the second network) and the observed concentration (data). The smaller the fitness, the better.
- An individual that cannot be simulated by SPICE is assigned a high penalty value of fitness (10^8)

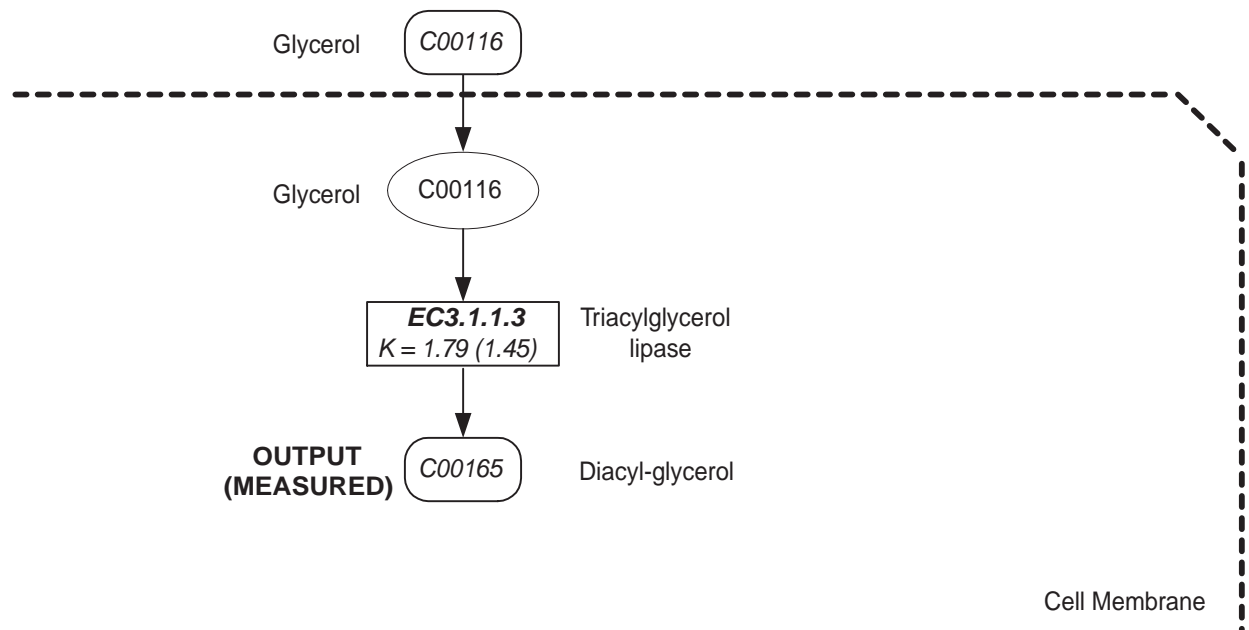
CONTROL PARAMETERS FOR THE RUN

- **The population size, M , is 100,000.**
- **A maximum size of 500 points (for functions and terminals) was established for each result-producing branch.**

RESULTS PHOSPHOLIPID CYCLE

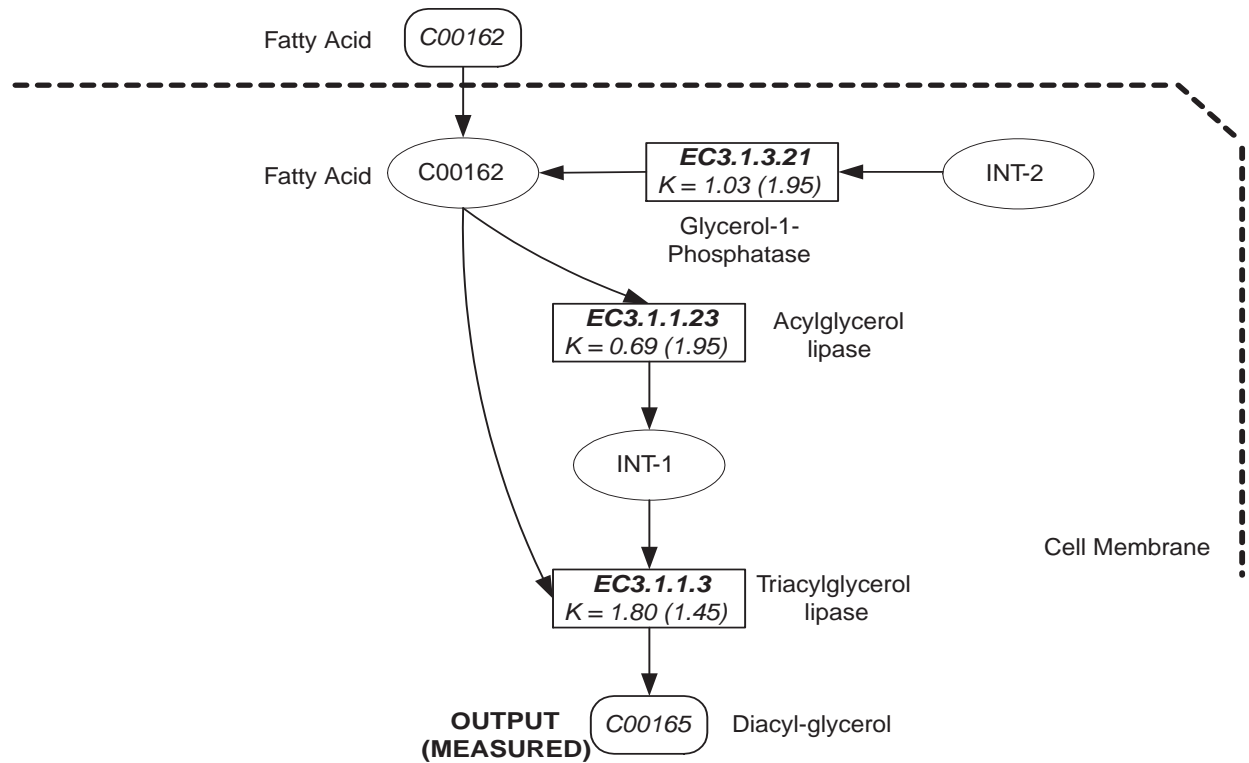
- The fitness of the median individual from the population at generation 0 is 297.3. This individual scores 17 hits (out of 270)

MEDIAN INDIVIDUAL OF GEN 0



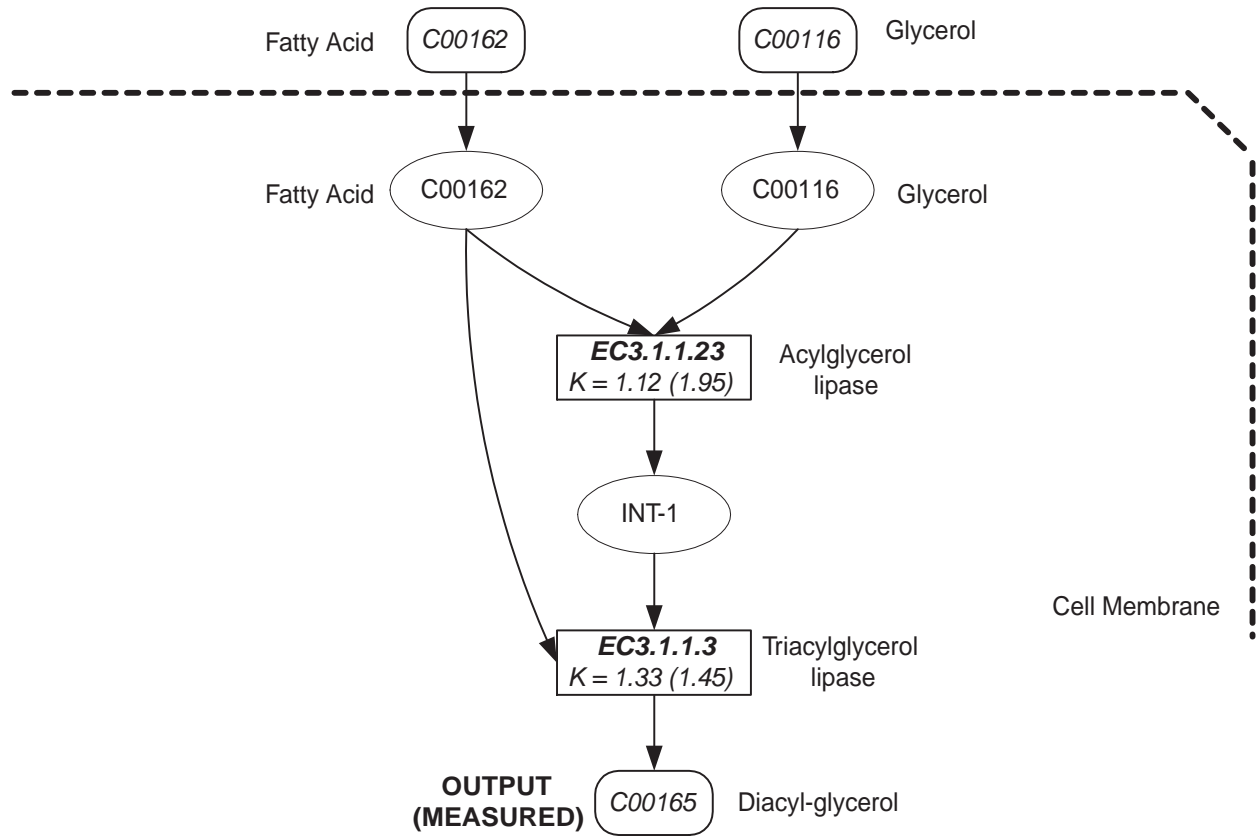
RESULTS PHOSPHOLIPID CYCLE

BEST OF GENERATION 0



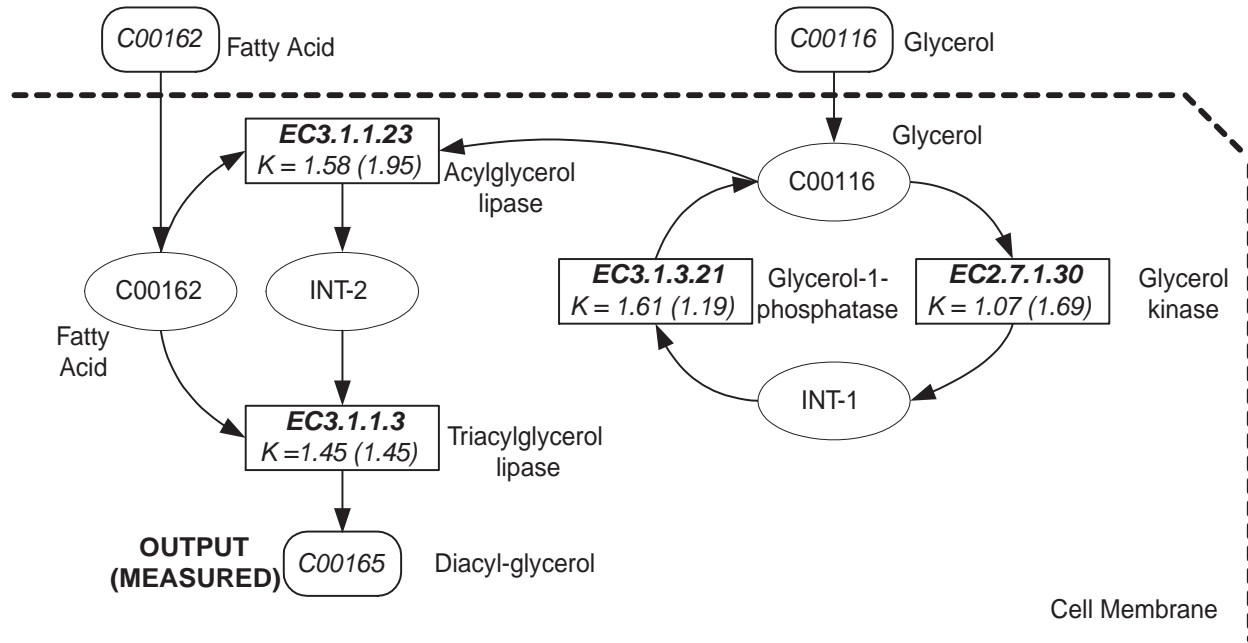
RESULTS PHOSPHOLIPID CYCLE

BEST OF GENERATION 10



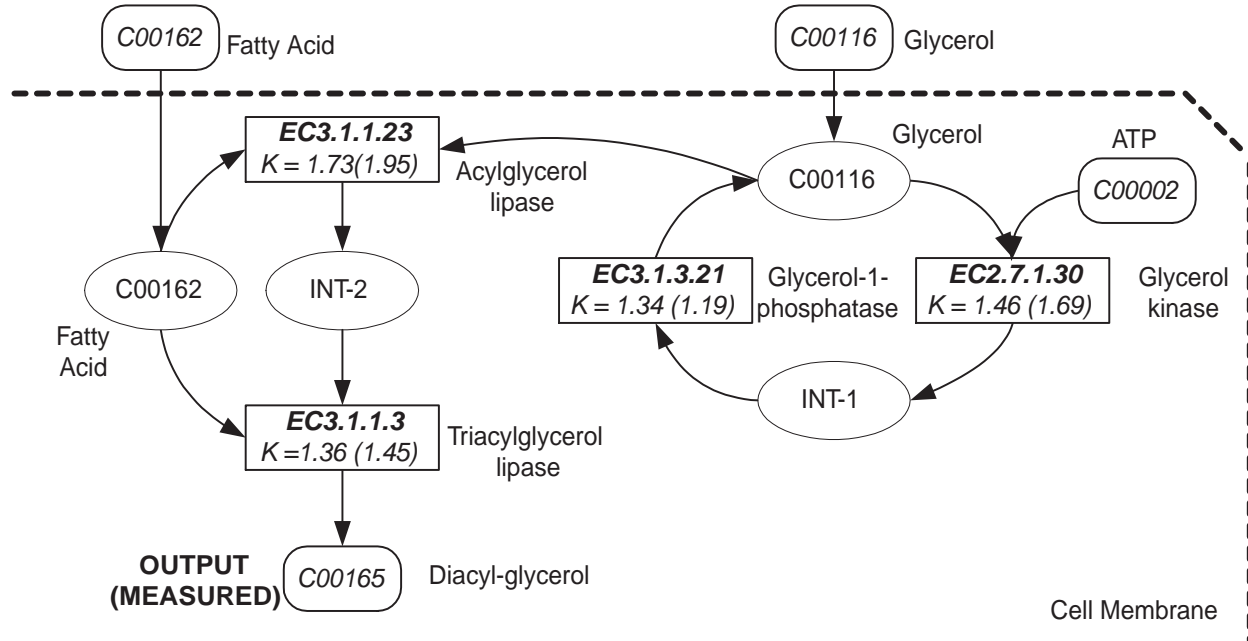
RESULTS PHOSPHOLIPID CYCLE

BEST OF GENERATION 25



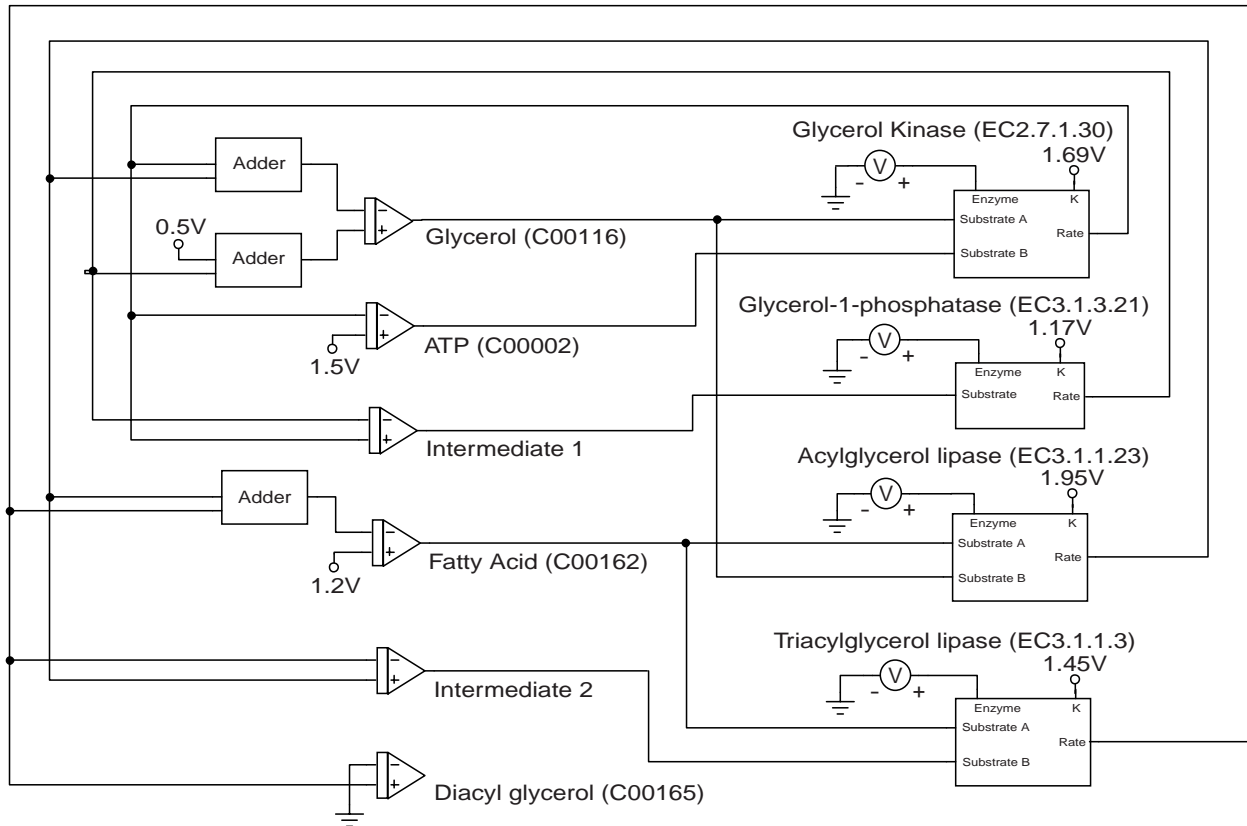
RESULTS PHOSPHOLIPID CYCLE

BEST OF GENERATION 120



RESULTS PHOSPHOLIPID CYCLE

ELECTRICAL CIRCUIT FOR THE BEST-OF-RUN INDIVIDUAL FROM GENERATION 225



RESULTS PHOSPHOLIPID CYCLE

- **Rate of production of the network's final product, diacyl-glycerol (C00165)**

$$\frac{d[C00165]}{dt} = 1.45[C00162][INT_2][EC3.1.1.3] \cdot$$

- **Rate of production and consumption of the intermediate substance INT_2**

$$\frac{d[INT_2]}{dt} = 1.95[C00162][C00116][EC3.1.1.23] - 1.45[C00162][INT_2][EC3.1.1.3] \cdot$$

- **Rate of production and consumption of the intermediate substance INT_1 in the internal feedback loop**

$$\frac{d[INT_1]}{dt} = 1.69[C00116][C00002][EC2.7.1.30] - 1.17[INT_1][EC3.1.3.21] \cdot$$

- **Rate of supply and consumption of ATP (C00002)**

$$\frac{d[ATP]}{dt} = 1.5 - 1.69[C00116][C00002][EC2.7.1.30]$$

- **Rate of supply and consumption of fatty acid (C00162) in the best-of-run network**

$$\frac{d[C00162]}{dt} = 1.2 - 1.95[C00162][C00116][EC3.1.1.23] - 1.45[C00162][INT_2][EC3.1.1.3] \cdot$$

RESULTS PHOSPHOLIPID CYCLE

- **Rate of supply, consumption, and production of glycerol (C00116) in the best-of-run network**

$$\frac{d[C00116]}{dt} = 0.5 + 1.17[INT_1][EC3.1.3.21] - 1.69[C00116][C00002][EC2.7.1.30] - 1.95[C00162][C00116][EC3.1.1.23]$$

- **Internal feedback loop in which C00116 is both consumed and produced**

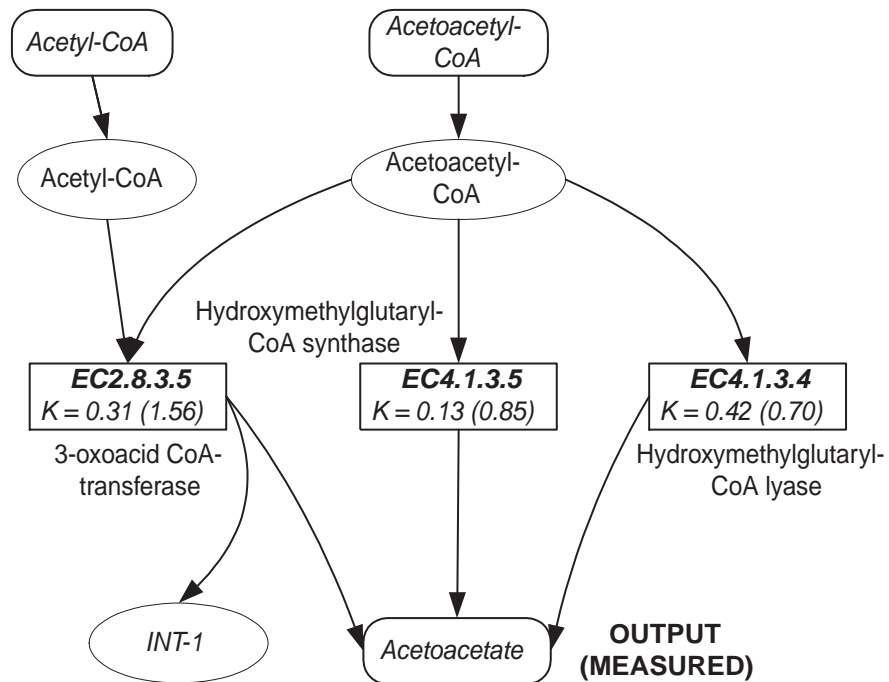
RESULTS PHOSPHOLIPID CYCLE

In summary, driven only by the time-domain concentration values of the final product C00165 (diacyl-glycerol), genetic programming created both the topology and sizing for an entire metabolic pathway whose time-domain behavior closely matches that of naturally occurring pathway, including

- the total number of reactions in the network,
- the number of substrate(s) consumed by each reaction,
- the number of product(s) produced by each reaction,
- an indication of which enzyme (if any) acts as a catalyst for each reaction,
- the pathways supplying the substrate(s) (either from external sources or other reactions in the network) to each reaction,
- the pathways dispersing each reaction's product(s) (either to other reactions or external outputs),
- the number of intermediate substances in the network,
- emergent topological features such as
 - internal feedback loops,
 - bifurcation points,
 - accumulation points, and
- numerical rates (sizing) for all reactions.

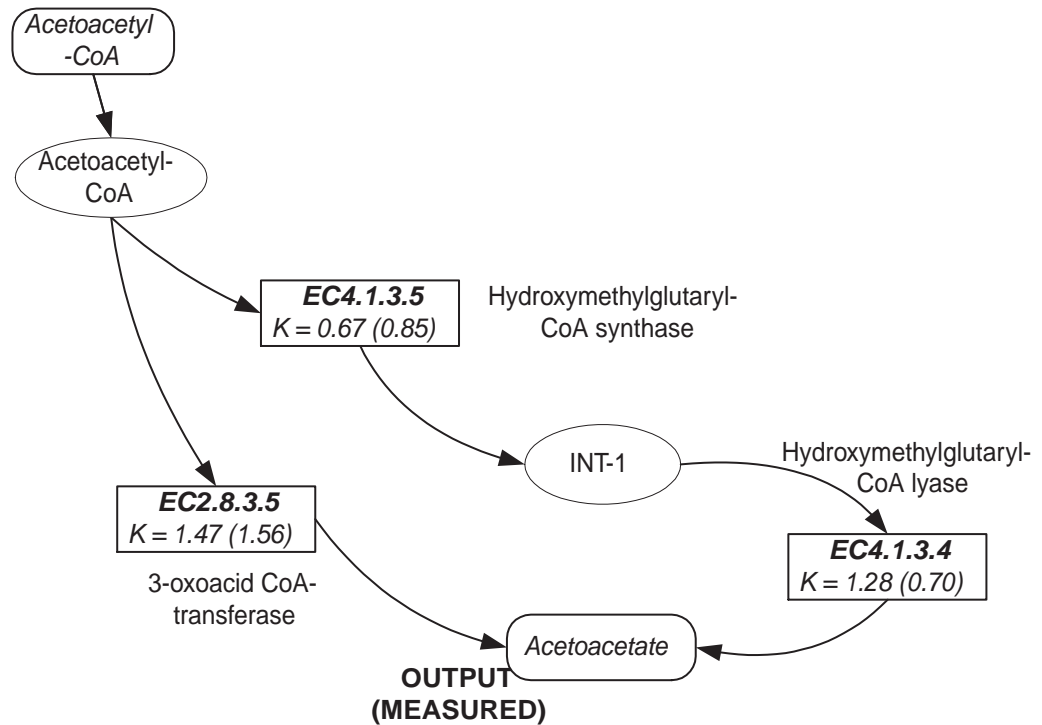
RESULTS — SYNTHESIS AND DEGRADATION OF KETONE BODIES

ONE INDIVIDUAL FROM GENERATION 0 WITH SEVERAL NOTEWORTHY TOPOLOGICAL FEATURES



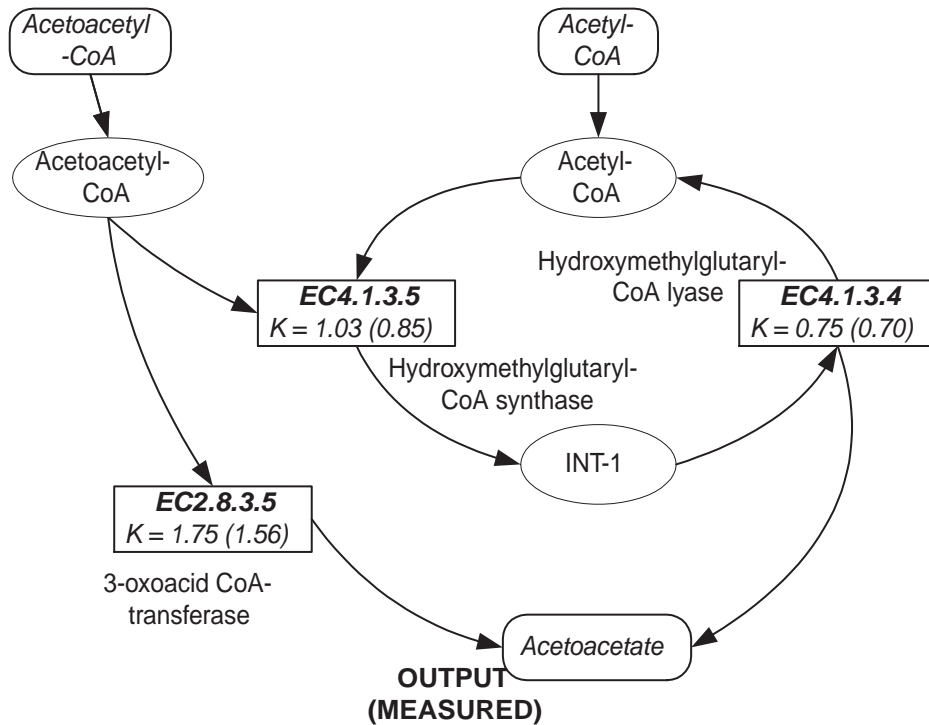
RESULTS — SYNTHESIS AND DEGRADATION OF KETONE BODIES

BEST NETWORK OF GENERATION 0



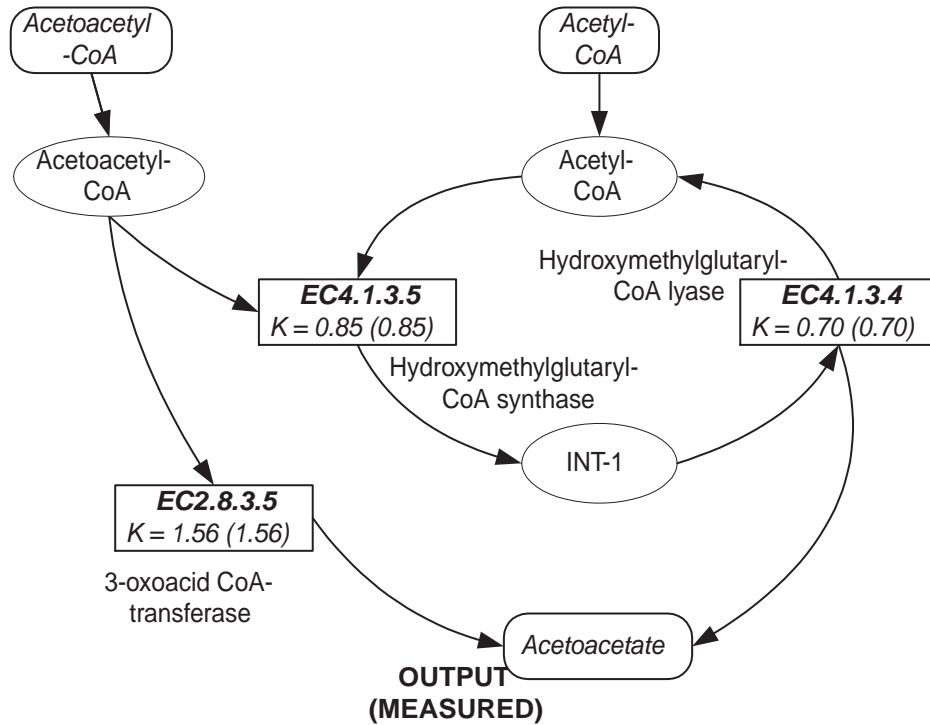
RESULTS — SYNTHESIS AND DEGRADATION OF KETONE BODIES

BEST NETWORK OF GENERATION 5



RESULTS — SYNTHESIS AND DEGRADATION OF KETONE BODIES

BEST-OF-RUN NETWORK OF GENERATION 97

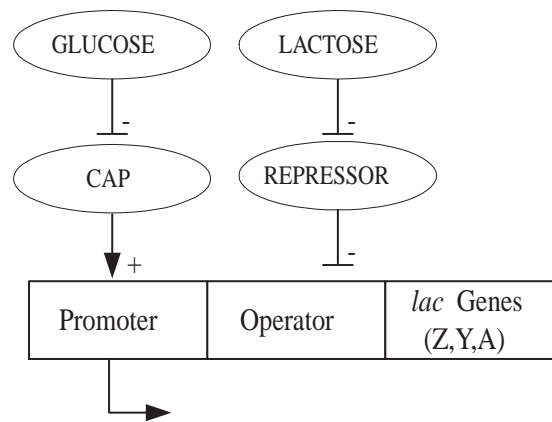


FUTURE WORK

- **Improved Program Tree Representation**
- **Multiplication and Division Functions**
- **Null Enzyme**
- **Minimum Amount of Data Needed**
- **Opportunities to Use Knowledge**
- **Designing Alternative Metabolisms**

GENETIC NETWORKS

GENETIC NETWORK FOR *LAC* OPERON



- This is a schematic representation of a genetic network for the expression level of the *lac* operon (composed of the Z, Y, and A genes).

GENETIC NETWORKS

- **The lac operon is a basic control circuit present in many simple organisms, including *Escheria coli*. The metabolism of lactose requires permease and b-galactosidase (encoded by the Z and Y genes, respectively).**
- **The permease is involved in the transport of lactose into the cell, while the b-galactosidase is involved in cleaving the lactose molecule into glucose and galactose.**
- **The purpose of this control circuit is to only express the proteins that metabolize lactose when glucose (the preferred source of energy) is scarce and lactose is abundant.**
- **The two regulatory proteins CAP and REPRESSOR are involved in regulating the expression of the Z and Y genes. The genetic network involves two proteins (REPRESSOR or CAP) and two substances (GLUCOSE or LACTOSE).**

GENETIC NETWORKS

- **The actual performance of the genetic network is determined by the expression levels of the two genes and the concentrations of the two substances in relation to threshold values.**
- **These threshold values serve as numerical parameters of certain conditional and comparative functions.**

GENETIC NETWORKS

- The logic underlying the genetic network for the *lac* operon can be succinctly written in C-style pseudo code. The numerical value returned by this program is the expression level of the *lac* operon (`LAC_mRNA_LEVEL`).

```
if(LACTOSE_LEVEL >= LACTOSE_THRESHOLD)
{
    if(GLUCOSE_LEVEL >= GLUCOSE_THRESHOLD)
    {
        LAC_mRNA_LEVEL = low;
    }
    else
    {
        if (CAP_LEVEL >= CAP_THRESHOLD)
        {
            LAC_mRNA_LEVEL = high;
        }
        else
        {
            LAC_mRNA_LEVEL = low;
        }
    }
}
else
{
    if(REPRESSOR_LEVEL >= REPRESSOR_THRESHOLD)
    {
        LAC_mRNA_LEVEL = 0;
    }
    else
    {
        LAC_mRNA_LEVEL = low;
    }
}
```

GENETIC NETWORKS

•The goal is to automatically create (reverse engineer) *both* a topological arrangement of conditional and comparative functions and all necessary numerical parameters that represent the expression level of the *lac* operon as measured by its mRNA. In other words, we seek to automatically create logic that is equivalent to that shown above using time-domain data for the expression levels of the two genes and the concentrations of the two substances.

GENETIC NETWORKS

REPRESENTATION OF GENETIC NETWORKS AS COMPUTER PROGRAMS

- **Each program tree represents the logic of a genetic network. A program tree is a composition of functions from the function set and terminals from the terminal set and contains**
 - internal nodes representing conditional and comparative functions,
 - external points (leaves) representing expression levels of various genes, and
 - external points representing concentration of substances.
- **The value returned by the result-producing branch of the program tree is the expression level of the *lac* operation (called `LAC_mRNA_LEVEL` in the C-style pseudo code above).**

GENETIC NETWORKS

REPERTOIRE OF FUNCTIONS

- **The three-argument `IF` function returns the results of evaluating its third argument (the "else" clause) if its first argument is `FALSE`, but returns the results of evaluating its second argument (the "then" clause) if its first argument is `TRUE`.**
- **The two-argument `<` comparative function returns a value of `TRUE` if its first argument is less than its second argument, but otherwise `FALSE`.**
- **The two-argument `>` comparative function performs the opposite function.**

GENETIC NETWORKS

REPERTOIRE OF TERMINALS

- The terminals **GLUCOSE_LEVEL** and **LACTOSE_LEVEL** represent substances.
- The terminals **REPRESSOR_LEVEL** and **CAP_LEVEL** represent expression levels of genes.

CONSTRAINED SYNTACTIC STRUCTURE

- The trees are constructed in accordance with a constrained syntactic structure. The entire program tree returns a floating-point number. The first argument of an **IF** function must be a comparative function (< or >). The two arguments of a comparative function must be a terminal. The second and third arguments of an **IF** function may be another **IF** function or a perturbable numerical value.

GENETIC NETWORKS

PREPARATORY STEPS

PROGRAM ARCHITECTURE

Each program tree has one result-producing branch.

FUNCTION SET

The function set is

$$\mathbf{F} = \{\mathbf{IF}, <, >\}$$

with arity of three, two, and two, respectively.

GENETIC NETWORKS

PREPARATORY STEPS

TERMINAL SET

In this problem, the numerical value(s) are established by value-setting subtree containing a single perturbable numerical value. These numerical values will serve as the thresholds in the overall logic of the evolved program.

- The terminal set, T_{vss} , for the value-setting subtrees is

$$T_{vss} = \{\mathcal{R}_p\},$$

where \mathcal{R}_p denotes a perturbable numerical value.

- The terminal set for all other parts of the program trees is

$$T = \{\text{GLUCOSE_LEVEL}, \text{LACTOSE_LEVEL}, \text{REPRESSOR_LEVEL}, \text{CAP_LEVEL}, \mathcal{R}\}.$$

GENETIC NETWORKS

PREPARATORY STEPS

FITNESS MEASURE

- **Each individual genetic network is exposed to four time-domain scenarios representing the concentrations of substances (GLUCOSE_LEVEL or LACTOSE_LEVEL) and expression values of genes (REPRESSOR_LEVEL or CAP_LEVEL) over 20 time steps (except that there are only 19 time steps in the first scenario since time $t = 0$ is ignored).**

GENETIC NETWORKS

PREPARATORY STEPS

FITNESS MEASURE

- **The first of the four fitness cases is based on a high level (10) of `GLUCOSE_LEVEL` and a low level (0) of `LACTOSE_LEVEL`. In this context the network is exposed, during the 20 time steps, to all four combinations of high and low values of `CAP_LEVEL` and `REPRESSOR_LEVEL`. Broadly speaking, the expression level of `CAP_LEVEL` initially rises. While `CAP_LEVEL` is steady, `REPRESSOR_LEVEL` begins to rise. When `REPRESSOR_LEVEL` reaches its peak, `CAP` begins to fall.**
- **The second fitness case is based on a high level (10) of `GLUCOSE_LEVEL` and a high level (10) of `LACTOSE_LEVEL`.**

GENETIC NETWORKS

PREPARATORY STEPS

FITNESS MEASURE

- **The third fitness case is based on a low level (0) of `GLUCOSE_LEVEL` and a high level (10) of `LACTOSE_LEVEL`.**
- **The fourth fitness case is based on a low level (0) of `GLUCOSE_LEVEL` and a low level (0) of `LACTOSE_LEVEL`.**
- **Fitness is the sum, over the 79 fitness cases, of the absolute weighted value of the difference between the value returned by the result-producing branch and the observed expression level of the *lac* operon (as measured by mRNA). If the value returned by the result-producing branch is within 5% of the observed expression-level data, the weight is 1.0; otherwise it is 10. The smaller the fitness, the better.**

GENETIC NETWORKS

PREPARATORY STEPS

FITNESS MEASURE

- **The number of hits is defined as the number of fitness cases (time steps 1 to 79) for which the difference is within 5% of the correct value.**

GENETIC NETWORKS

PREPARATORY STEPS

CONTROL PARAMETERS FOR THE RUN

The population size, M , is 10,000.

GENETIC NETWORKS

BEST INDIVIDUAL FROM GENERATION

93

```
(IF (< LACTOSE_LEVEL 9.139 ) (IF (< REPRESSOR_LEVEL
6.270 ) (IF (> GLUCOSE_LEVEL 5.491 ) 2.02 (IF (<
CAP_LEVEL 0.639 ) 2.033 (IF (< CAP_LEVEL 4.858 ) (IF (>
LACTOSE_LEVEL 2.511 ) (IF (> CAP_LEVEL 7.807 ) 5.586
(IF (> LACTOSE_LEVEL 2.114 ) 1.978 2.137 ) ) 0.0 ) (IF
(> REPRESSOR_LEVEL 4.015 ) 0.036 (IF (< GLUCOSE_LEVEL
5.128 ) 10.0 (IF (< REPRESSOR_LEVEL 4.268 ) 2.022 9.122
) ) ) ) ) (IF (> CAP_LEVEL 0.842 ) 0.0 5.97 ) ) (IF
(< CAP_LEVEL 1.769 ) 2.022 (IF (< GLUCOSE_LEVEL 2.382 )
(IF (> LACTOSE_LEVEL 1.256 ) (IF (> LACTOSE_LEVEL 1.933
) (IF (> GLUCOSE_LEVEL 2.022 ) (IF (< GLUCOSE_LEVEL
5.183 ) 6.323 (IF (> CAP_LEVEL 1.208 ) 9.713 0.842 ) )
10.0 ) (IF (> GLUCOSE_LEVEL 6.270 ) 2.109 ) 1.965 ) )
0.665 ) 1.982 ) ) )
```

GENETIC NETWORKS

BEST INDIVIDUAL FROM GENERATION 93 REWRITTEN IN C-STYLE PSEUDO CODE

```
if(LACTOSE_LEVEL < 9.139)
{
    if(REPRESSOR_LEVEL < 6.270)
    {
        LAC_mRNA_LEVEL = 2.022;
    }
    else
    {
        LAC_mRNA_LEVEL = 0.0;
    }
}
else
{
    if(CAP_LEVEL < 1.769)
    {
        LAC_mRNA_LEVEL = 2.022;
    }
    else
    {
        if(GLUCOSE_LEVEL < 2.382)
        {
            LAC_mRNA_LEVEL = 10.0;
        }
        else
        {
            LAC_mRNA_LEVEL = 1.982;
        }
    }
}
```