Evolution of Complexity in Biochemical Pathways

Jay Kim

Submitted to the Vassar College Department of Biology

Undergraduate Thesis
Academic Year: 2016 - 2017

First Reader: Professor John Long
Second Reader: Professor Leah Bendavid

Acknowledgements

This thesis was completed as a result of the endless kindness and support of those around me.

First and foremost, I am forever indebted to Professor John Long for his unconditional support and belief in this project and me. Without his truly incredible level of patience and belief in me, there is no doubt that this thesis would either not have been completed, or would it not remotely be of any acceptable quality. Professor Long's guidance has always been illuminating and understanding, even in the harshest of circumstances, and has been a light in this journey of mine.

I would also like to thank Professor Leah Bendavid for providing invaluable modeling advice as well as a unique perspective on developing biological models and their real-world applications.

While it would be futile to list all of my friends that have supported me in this thesis and graciously listened to my frustration and provided feedback, I would like to give special notice to Kentaro Kaneki, Mariah Marshall, and Wendy Liu.

Finally, I would like to also thank my family for their constant support. They have always been around me since day one, and have always believed in me.

Table of Contents

Abstract

Due to recent advents in computing power and software, agent-based models (ABM) are becoming a popular method to simulate biological phenomena. However, much remains to be seen about its application on cellular processes at a mesoscopic level due to a lack of experimental data. The difficulty of parameter estimation, particularly for multivariate ABM models, also makes modeling multiple biological phenomena within a single model difficult. This project attempts to model a proto-cell on a mesoscopic level, using molecules as agents. I employed a novel approach to isolate a biochemical pathway of interest within a simulation without neglecting the effects of side reactions and flux in the proto-cell. I observed that environmental stress on membrane formation pathways, rate of gene expression, flux, and side-reactions may have non-linear impacts on the development of the proto-cell. Furthermore, the model used in this project demonstrates the effects of paralogous genes, gain-of-function mutations, and mutations that affect an enzyme's capability on a biochemical pathway. Here, paralogous genes and gain-of-function mutations allow recovery of the proto-cell from otherwise fatal environmental conditions, whereas the greater variance from a greater likelihood and strength of mutations often kill off a large degree of the proto-cells.

Introduction

The origin and development of the complexity and diversity of life is a question of much interest to scientists. This is because a deeper understanding of the necessary conditions and characteristics of systems that allow life to survive and evolve may advance the development of evolutionary systems. To do so, a popular method of investigating the characteristics of these evolutionary systems is the use of *in silico* models [1, 2], due to the difficulty, lack of knowledge, and the inability to collect data from reproducing early biological and evolutionary systems. Two approaches have been used, equation-based modeling (EBM) and agent-based modeling (ABM). The former is characterized by the usage of equations to simulate macroscopic behaviors and may be favored for its low computational costs and the prevalence of known macro-scale equations. The latter is relatively new and is characterized by its high computational cost and the simulation of autonomous agents that act according to set rules, which are based on the agent's mesoscopic behaviors. For the reasons above, the simulation of virtual organisms have been dominated by EBM. However, with the advent of increased computational power, ABMs are becoming more popular due to its lower levels of abstraction. These lower levels of abstraction may allow models to become more accurate to their real-life counterparts as they may relax certain assumptions made from EBMs, which may be valuable in the context of certain experiments such as simulation of the immune system [2-4] and may explain certain behaviors that are lost in the gap between macroscopic and microscopic rules. Current ABMs within this field have explored intracellular kinetics or have been abstracted into modeling the behaviors of whole

organisms but have avoided an agent-based model of a cell on a molecular level due to the lack of requisite information.

Furthermore, the relationship between intracellular reaction kinetics, evolution, and environmental changes is not directly well understood. Within *in vivo* or *in vitro* experiments, the data is difficult to obtain. For *in silico* experiments, either the high level of abstraction used in equation-based modeling or the lack of information on mesoscopic behaviors renders model building difficult [5]. This lack of information comes as a direct result of the usage of macro-scale system measurements to define the behavior of a collection of molecules, rather than identifying the particular parameters that contribute to these macro-scale system measurements, due to the appearance and observations of macro-scale systemic behaviors in real-world applications of chemistry.

As a result, simulations of cellular activity and behaviors have been abstracted either to an organismal level, utilized macro-scale behaviors to define molecular activity, or have focused on particular cell processes [6]. Well-known simulations within the field, such as the whole-cell simulation of *Mycoplasma genitalium,* and Turing-complete cellular automata are generally constrained by such limitations and therefore are difficult to generalize into unexpected conditions [7, 8]. Of the biological ABMs that currently exist, there is a strong focus in modeling only a single level of agent-agent interactions and abstracting the complex hierarchies that exist across various levels of interactions. That is, the agents in such models act according to a set of defined rules, but the underlying phenomena that allow these defined rules to occur and thus the agent-agent interactions, are ignored [6, 9]. As a result, current ABMs may model enzyme kinetics or the reproduction of bacteria in a given environment, but never a hybrid approach that

directly models how enzyme kinetics may lead to the reproduction of such bacteria [4]. Furthermore, the high computational cost of simulating the number of molecules that exist within a cell makes the creation of a cell on a mesoscopic level immensely difficult [6, 9]. Thus, there exists a divide between macroscopic and mesoscopic levels.

Further complicating simulations of life is the difficulty in identifying all of the pathways and interactions necessary for life. Although minimal-life cell models, experimental creations with the minimal amount of genes necessary to satisfy the definitions of life, have been created, and in such minimal-life models the requisite genes for life have been recognized, the particular interplay amongst all of the various genes, as well as the non-genetic elements necessary for life are well beyond our current grasp of understanding [10]. Despite the various definitions for the demarcating point between living and non-living [11], amongst biologists these conditions are generally agreed to be a stably self-replicating, self-compartmentalized membrane that is capable of undergoing metabolism and can undergo some form of hereditability with variance [12]. As such, much attention has been focused on creating proto-cells, precursors to life but deficient to satisfying the conditions of a biological organism. Experimental attempts at creating life *de novo* have not yet yielded a universally agreed-upon successful proto-cell, although Bahadur's Jeewanu model remains a possible candidate [13]. Amongst theoretical models, Ganti's chemoton model, Eigen and Schuster's hypercycle theory, Gilbert's RNA world, Wächtershäuser's metabolism first, and Luisi's lipid world theory all create an intriguing glimpse into the possible prebiotic conditions required for life to emerge [12].

This thesis hopes to create a model to reconcile the levels of abstraction present in macro-scale and meso-scale modeling to progress the understanding of evolution with

regards to pre-biotic biochemical pathways. In doing so, it becomes possible to discern some of the requisite characteristics of the units of a system for complexity and emergence to occur within self-replicating organisms.

For this project, I propose a mesoscopic agent-based model of a proto-cell, complete with a functional genome, to observe how mutations to enzymatic agents and how particular attributes to the proto-cell (such as mutation rate and prevalence of side-reactions) can affect the evolution of a biochemical pathway. Several simplifying assumptions were made to compensate for the necessary computational cost and complexity in modeling early life. To retain real-world significance, the biochemical pathway constructed for this simulation was based off of Ganti's chemoton model [12] and Eigen and Schuster's hypercycle theory [14], and the proto-cell is bounded, self-replicating, and capable of producing enzymes with slight variance so that the proto-cell resembles possible early life after the formation of self-compartmentalized membranes with metabolic cycles or functional RNA.

Chapter 1

Model Design

1.1 Agent-based model approach

For this particular project, an agent-based approach was used. An agent-based approach was preferred compared to the more common equation-based approach for several reasons:

1. Agent-based Models allows a direct understanding of the interactions required to simulate complex phenomena (the transition between micro level behaviors to macro level behaviors) at a mesoscopic level. Analysis and collection of specific behaviors/dynamics are possible in a way that is difficult to obtain under equation-based modeling.

2. An equation-based approach relies on key assumptions that may not be viable under the

conditions being tested. That is, dynamic environments or interactions can be tested in a way that may be difficult under some equations [5].

3. Certain behaviors do not need to be explicitly modeled; these behaviors are a result of basic properties of the agents. For example, while examining the effects of a heterogeneous collection of agents versus a homogeneous collection of agents may require different equations in an EBM to model factors such as viscosity, an ABM implicitly accounts for the effect on viscosity if the appropriate parameters are given.

Although there are certain limitations with using these ABM, particularly with the amount of available techniques viable for cellular modeling and the computational cost of simulating what would be considered in the real world to be minute quantities of molecules, these limitations are considered a matter of hardware constraints and the recency of the ABM field rather than fundamental flaws within the ABM approach [5]. Thus, by modeling a proto-cell via an ABM approach, techniques used here may be applicable for future developments of more realistic and practical designs of proto-cells.

Furthermore, ABMs have a particular benefit of being user-friendly and beginner-friendly, an advantage that most equation-based models and paradigms such as Virtual

Cell or E-Cell do not offer. Languages such as NetLogo or Swarm are intentionally created to have a low barrier of entry by simplifying entire procedures into keywords while maintaining their flexibility and usability for various models. While such languages also have certain drawbacks due to the simplification of various procedures and thus the inability to separate particular processes or optimize the code, they remain a practical option for undergraduates interested in modeling bottom-up systems such as the ABMs [2, 15].

1.2 NetLogo

A mesoscopic agent-based model was created using NetLogo 5.3.1, an open source multi-agent programming environment written on the Java virtual machine. NetLogo uses a highly accessible programming language and its usage for agent-based models is well documented [2]. Its popularity as an environment for ABMs are due to its intentional simplistic design, created to mirror the philosophy of the Logo programming language "low threshold and no ceiling" [2]. Agents are identified in the form of "turtles" while the spatial area of the simulation is defined in terms of "patches". There exists an "observer" that is capable of manipulating the world that these agents operate in, but are incapable of directly affecting the behaviors through keywords or commands. Rather, the agency is conferred upon the agents that react according to the world's conditions. As a result, there are three types of variables within the NetLogo environment:

1. Global variables which are utilized by the observer and affect the world,

2. Turtle-own variables which are utilized by the turtles and may be unique per turtle
3. Patch-own variables, which are similarly utilized by patches and may also be unique per patch.

Since global variables affect the world that these turtles operate in, all turtles must act and react according to the globally held values of these variables. In contrast, turtle-own and patch-own variables may be unique per turtle.

To operate, this particular model characterizes four different classes of agents, enzymes, substrates/metabolites, products, and a genome. Every agent in the initial conditions of the model is randomly positioned and given a random orientation. The entire simulated space is unbounded; the simulated environment may be thought of as the inside of a torus, where agents can move endlessly in one direction and return to its original position. The entire world (simulated space) is considered to be the proto-cell, and consists of 200x200 patches (x- and y- coordinate of the environment). Each patch is 0.001 μm, for a $2\mu m$ x $2\mu m$ proto-cell, a common size for bacteria [16]. For simplifying purposes, this proto-cell environment exists in a 2-D world. Each time-step is equivalent to 1 real world second. The viscosity of the proto-cell is approximated as 0.0011 kPa [17].

To create and start a model, NetLogo requires a "setup" command that defines the initial conditions and rules of the simulation. After setup is complete, NetLogo requires an alternative procedure called "go" to run the simulation (see Figure 1). Each iteration of the go procedure increases the time-step of the model by 1 unit, or tick. Most rules and commands are nested in some manner within the go command and may be triggered either upon a particular control flow statement such as an "ifelse" command or will be

constantly called every iteration of the go procedure. While NetLogo simplifies a vast amount of ABM programming through specific keywords, the high-level nature of NetLogo limits the amount of options one has in modeling. As a result, one is able to extend the number of keywords through the addition of external "extensions", which are written in Java and supplement the NetLogo software.

```
to go
  ;; *Place end conditions here as a conditional*
  ask turtles
  [check-for-collision
  diffusion
  move
  if not (breed = genome) [set next-gen-tag 0]

  ;; Add rxn-functions here. For speed, write an if function
  ;; specifying the breed and rxn-reactant conditions, as shown below

  if breed = a-1-n and (rxn-reactant = "complex")
    [ask self [dissociate-a1-&-s 15 3 3]]
  if breed = a-1-n and (rxn-reactant = nobody)
  ;; If an agent has more than 1 possible reaction, add it inside the below bracket
  [a-1-n-&-sub-rxn]
```

Figure 1. Go procedure in NetLogo. First 16 lines of the go procedure. Full version of the go procedure found in the S1 of the supplementary materials.

To identify different agents within the simulation, the keyword "breeds" is used. Breeds are used as an identifier of a set of agents, and additional procedures are necessary for defining the actual behaviors and characteristics of these agents. To change turtle-own variables or to initiate certain behaviors, one must "ask" either the agent-set of all turtles within the simulation, turtles that fall under a specific condition, or the breed of the turtle, to execute the desired command.

NetLogo simulations are visually represented in the Interface tab of NetLogo. Here, one may observe the interactions between turtles and may create "buttons",

"sliders", "monitors", and a variety of other interface-specific tools to either visualize the changes occurring within the simulations or to change the value of the global variables during the simulation.

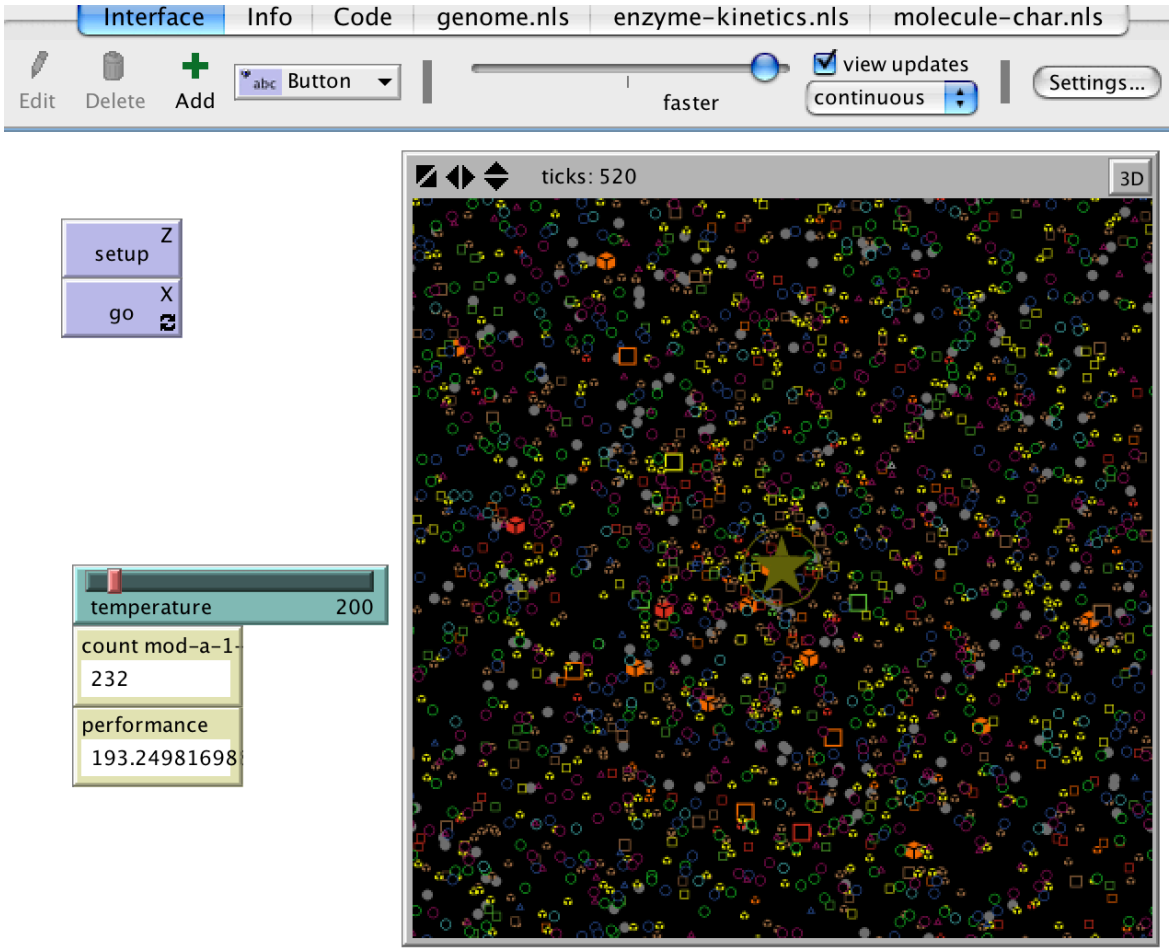


Figure 2. NetLogo Interface tab. One can observe the visual representation of the simulation in the large box on the right. Particular parameters may be adjusted through the slider and measurements may be observed during the simulation through the "monitors". The "setup" and "go" buttons are present on the upper left corner and trigger the "setup" and "go" procedures.

One useful feature of NetLogo is the implementation of the Behavior Space feature. Behavior Space allows one to run the model many times while recording desired results in a .csv file. In running the model, Behavior Space is also capable of multiple model runs in parallel to each other. That is, NetLogo is able to allow a version of the model to run a defined number of times, with the software running a smaller subset of these models in parallel to each other [2].

Issues exist however with the usage of Behavior Space, especially in terms of data collection and computational cost, thus limiting its practical effectiveness. For instance, allowing parallel runs distorts the data, requiring data wrangling techniques as all runs are recorded in the same .csv file and consequently, parallel runs lead to interleaved, out-of-order results. Furthermore, only global data and variables are able to be collected. To access the turtles-own variable, one must create his/her own data collection procedure, one that is fundamentally incompatible with particular aspects of the Behavior Space feature. An increase in the number of turtles within the simulation exponentially increases the amount of data collected by NetLogo, allowing data as large as 5 GB to be collected within a single run. Furthermore, to create this data collection procedure, one must open a file every time a model run is initiated, a procedural call that runs in conflict with the Behavior Space feature. Thus, the practicality of Behavior Space depends ultimately on its usage. If one seeks to only gather global data, then Behavior Space drastically reduces the time/effort necessary for collecting data. However, if turtle-own data must be collected, then alternative procedures are required.

For this thesis, such an alternative procedure was created through the following code.

```
;; opens file for the data to be saved into
to new-file
  let file user-new-file
  if is-string? file
  [
; if file-exists? file
;    [file-delete file]
    file-open file
    ;; Need to print the headers
    file-print csv:to-row
    (
      list "Ticks" "ID" "Compound name" "evo-tag" "generations" "performance-1"
      "performance-2" "performance" "reaction_total_type-1" "reaction_total_type-2"
      )
    write-to-file
  ]
end

;; obtains data from the turtles
to-report get-vals
  report
  (list ticks who compound-name evo-tag generations performance-1 performance-2
    performance
    rxn-1-counter rxn-2-counter
    )
end

;; allows turtle data to be written in the new file
to write-to-file

  ;; use SORT so the turtles print their data in order by who number,
  ;; rather than in random order
  foreach sort turtles [
    ask ? [
      file-print csv:to-row get-vals
    ]
  ]
  file-print ""  ;; blank line
end
```

Figure 3. Turtles-own data collection procedure.

To obtain turtles-own data, it is necessary to call for NetLogo to open a .csv file, which is what the new-file procedure does. The new-file procedure also delineates the turtles-own data into columns through the use of the “csv” extension. The get-vals procedure records the desired turtles-own data into a list and is called by the write-to-file procedure so that it applies to all turtles within the simulation, and so that these lists are

recorded within the .csv file.

1.3 Proto-cell development and pathways

The proto-cell created for this project was based off of Ganti's chemoton model [12]. The chemoton model is a construction of a series of biochemical pathways that fulfill what Ganti described as the minimal properties necessary for life. Here, Ganti states "… the fundamental unit (i.e. the minimal system) of biology must have some specific properties:

> It must function under the direction of a program
>
> It must reproduce itself
>
> It and its progeny must be separate from the environment " [12].

To satisfy these conditions, Ganti's chemoton model consists of three autocatalytic cycles, each having a unique and specific purpose. Based off of Eigen and Schuster's hypercycle theory of autocatalytic cycles [14], the cooperation and competition between these autocatalytic cycles allow the system containing these autocatalytic cycles to be robust and stable against a dynamic environment and also direct the expansion of these cycles into more complex pathways.

The first autocatalytic cycle is related to the metabolism of the system and is responsible for the continual reproduction of the components of all of the autocatalytic cycles present in the system. The second autocatalytic cycle deals with the formation of a compartmentalized membrane, which is dependent on the compounds created from the first autocatalytic system. The third autocatalytic system serves as an early form of a genome, capable of producing macromolecules through template polycondensation through the molecules created from the first autocatalytic cycle.

The likeliness of these autocatalytic cycles during the early stages of the formation of life is supported by later theoretical evidence from simulations. According to Hordjik [18] and Vasas et al. [19], if the food set, or the set of available molecules within a system, is sufficiently large, has a low level of complexity, and retains a moderate level of catalysis within the system, the production and introduction of new species to the system will likely initiate and sustain new autocatalytic cycles, based on the initial available food set. Thus, Ganti posits that the formation of a subset of these naturally occurring autocatalytic cycles provide the required ingredients for the creation of life.

When these autocatalytic cycles work in conjunction with each other in a spherical membrane, they can allow the system to self-replicate as a result of decreased osmotic pressure and destabilization of the system. Although the system produces membrane-

forming components at the same rate as other molecules according to Ganti's chemoton model (see Figure 1), the increase in the volume of the spherical system occurs at a faster rate than the increase in the surface of the sphere as a result of the 3-D nature of the system. Whereas the volume of a sphere is $V = \frac{4}{3}\pi r^3$, the surface area of a sphere is $S.A. = 4\pi r^2$. As a result of the difference in power of the radius, the system faces an overall decreased concentration as time increases, destabilizing the system's shape and decreasing the osmotic pressure within the system. According to Ganti, the instability and deformation of the system's membrane as a result of this decreased osmotic pressure inevitably results in the division of the system into two equal spherical systems, producing a self-replicating proto-cell [20].

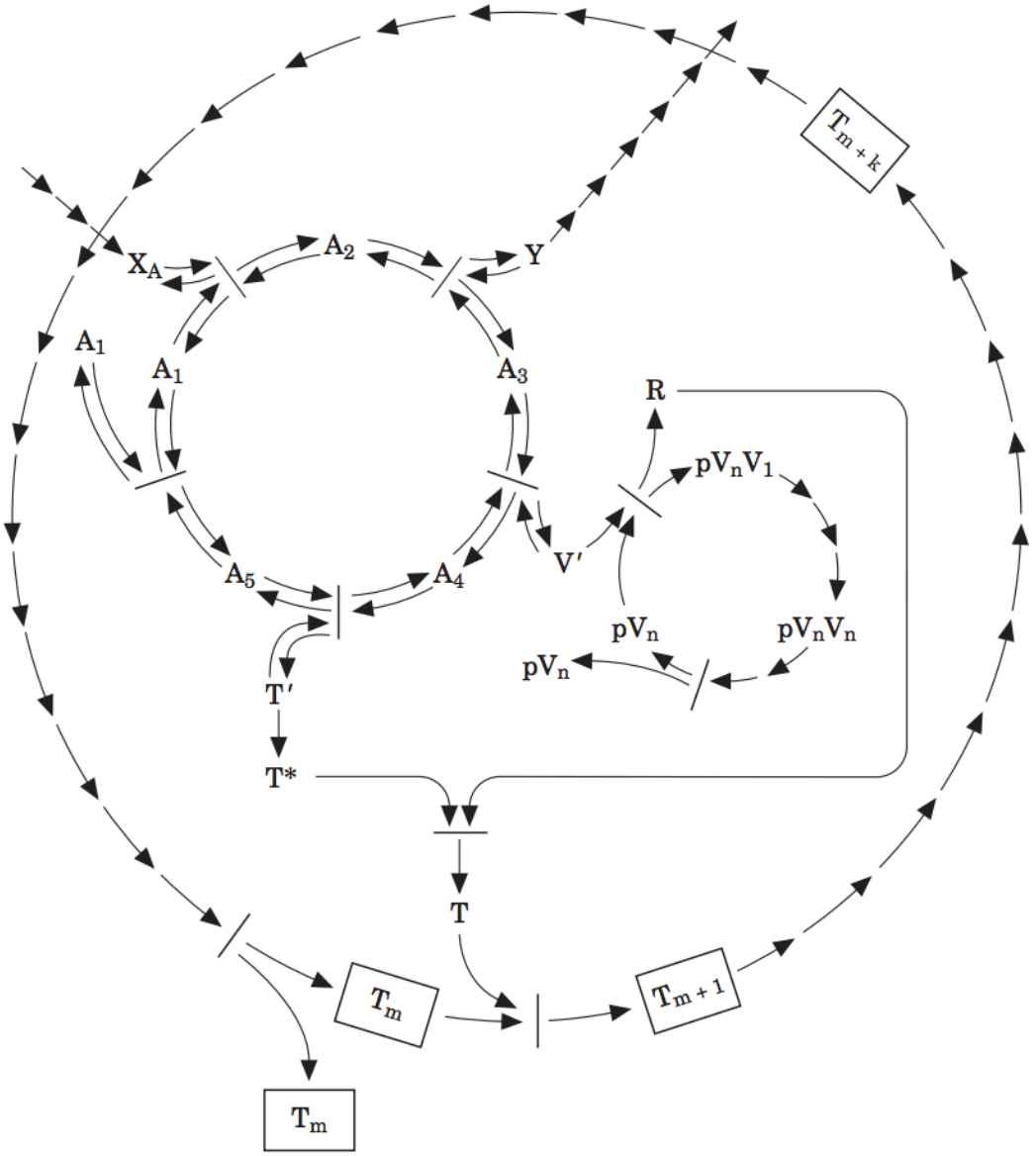


Figure 4. Ganti's chemoton model. Molecules represented by X_A are exogenous from the environment. Molecules of A_N are involved in the autocatalytic metabolic cycle, which creates compounds necessary for the membrane formation autocatalytic cycle, T_N, and the template polycondensation, or genome-related, autocatalytic cycle, pV_N. R represents a molecule created from the template polycondensation cycle that is necessary for

membrane formation. Similarly, Y represents a molecule from the metabolic cycle that is also necessary for the membrane formation cycle. Image from Ganti, 2002.

The particular model for this project examines a particular autocatalytic cycle from Ganti's chemoton model and expands upon it. The biochemical pathway selected for this project's model is situated amongst the membrane formation and replication of the proto-cell, where the genome-like autocatalytic cycle and the metabolic autocatalytic cycle produces the necessary components for the membrane formation autocatalytic cycle. Rather than assuming an autocatalytic cycle for the genome, which resembles what is believed to be the absolute beginnings of this self-sustaining system, this project positions itself in the later time period where a single genome-like component is responsible for the constant production of pre-biotic enzymes. In doing so, one can avoid some of the complications that are likely to have occurred during the absolute beginnings of the formation of the chemoton, where multiple genome-like autocatalytic cycles were competing against each other and limited in the pre-biotic enzymes they could produce [14, 21]. Furthermore, the project skips the early phases of the chemoton where there likely existed a divide between the enzymatic agents and the membrane formation cycles, and orients itself in the time period where the enzymatic agents are directly involved in membrane formation. Such a leap in the transition of the chemoton into early sustainable life is well justified by numerous experiments within the literature that suggest that although it is likely that genome-like did not initially create compartmentalized membranes, later synthesis from available pre-biotic organic materials allowed the genome-like structure/cycle to maintain direct control over membrane formation and

replication [22, 23]. This departure from the chemoton model also avoids some of the implicit assumptions created within the chemoton model. In Ganti's model, one can observe the critical role that the autocatalytic metabolic cycle plays within the functionality and propagation of the system. As a result, a metabolism-first approach is implicitly favored, as it assumes that the metabolic system serves as the precursor for all other aspects of the system. In contrast, a RNA-world approach suggests the development of the metabolic cycle from early genome-like structures, and so a dependency exists on the autocatalytic genome-like cycle rather than the metabolic cycle. For this thesis, we situate ourselves within either a RNA-world centered approach or within Ganti's metabolism-first approach, where later development and integration of the genome-like structure into the metabolic autocatalytic cycle has occurred.

Thus, I created the following autocatalytic membrane formation cycle and integrated the genome-like structure/cycle as constitutively expressing the enzymes necessary for the membrane formation cycle's integrity (see Figure 2). To explore evolutionary events within this model, functional mutations are permitted to occur probabilistically. Functional mutations are defined as mutations that alter the enzyme's behavior and operations through either gaining a new function or through creating a paralogous gene. While the former functional mutation merely alters the existing enzymes produced from the gene, the latter functional mutation creates a new gene with a similar but slightly altered enzyme that is able to be produced concurrently with the original gene. To characterize enzymes based on whether they have undergone a particular mutation or not, normal enzymes that

have not undergone mutations are given the name A-#-N enzymes, where # stands for a number from 1 – 5, the number of possible enzymes within the autocatalytic cycle. Enzymes that have undergone the functional mutation of a gain-of-function are given the name A-#-G enzyme to demonstrate the new gain-of-function. Similarly, new enzymes created from the paralagous gene are given the name of A-#-R enzyme, to show the relation of the paralgous gene to the original gene. These paralogous genes serve in parallel to the original gene. For example, in Figure 2, the A-1-N enzyme is responsible for the reaction $Substrate_1 + A - 1 - N \rightarrow Substrate_2$.

The A-1-R enzyme is responsible for the parallel reaction in the expanded topology:

$$Modified\ Substrate_1 + A - 1 - R \rightarrow Modified\ Substrate_2,$$

with the difference in that it reacts with the slightly modified version of the original substrate. The occurrence of a functional mutation on a particular enzyme is independent of the occurrence of a functional mutation of another particular enzyme. However, the gain-of-function functional mutation is deemed to be a pre-requisite for the paralogous gene functional mutation to occur.

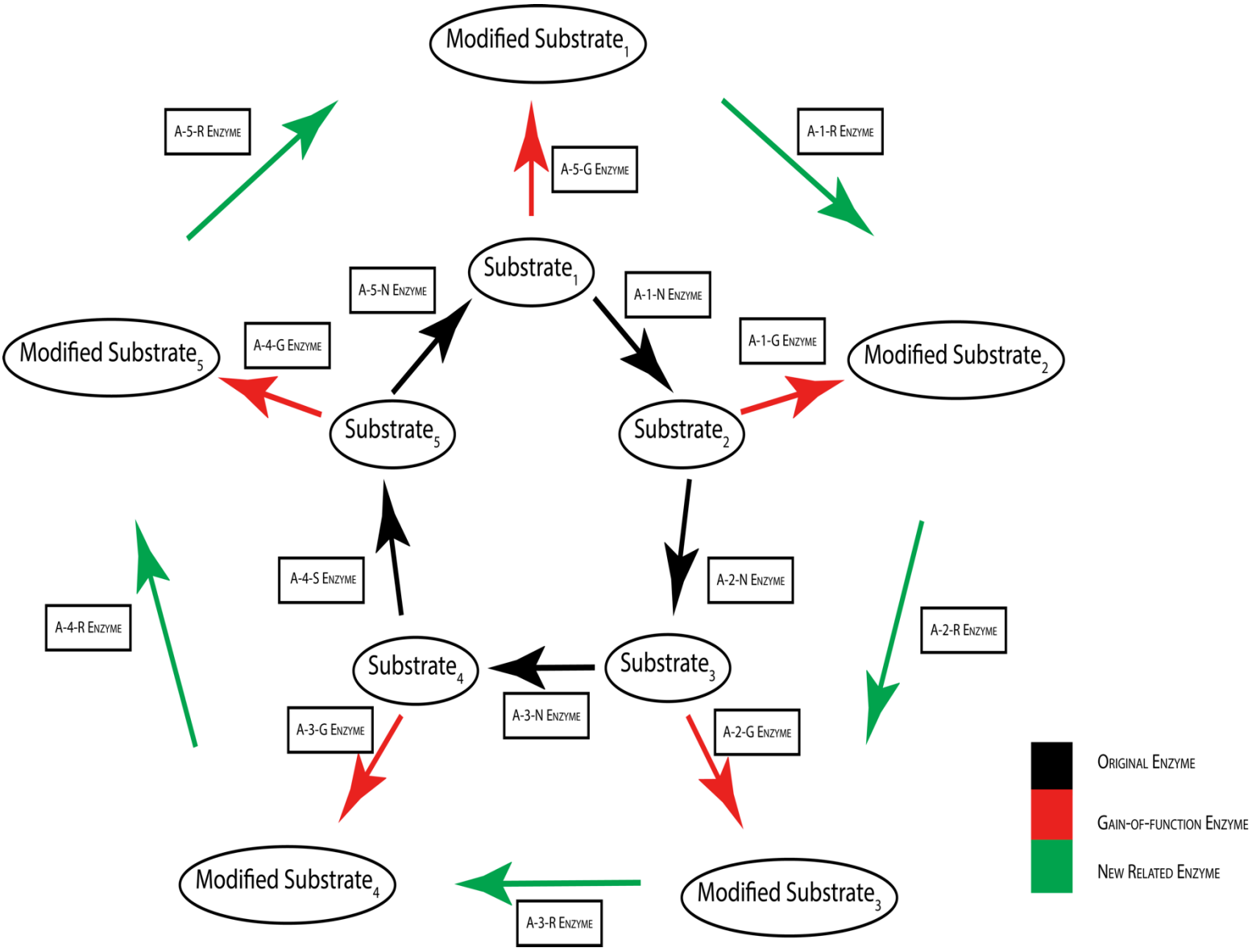


Figure 5. Topology of selected biochemical pathway for thesis model. Enzymes in the initial pathway will probabilistically evolve and react with both the original substrate and a new substrate. These gain-of-function enzymes may also trigger the duplication of the gene. The duplicated gene will produce an enzyme similar to the original enzyme, but bind to a different substrate (paralogous gene).

To allow the genome to retain memory of the functional mutation events and produce the gain-of-function or paralogous genes, a tag is introduced whenever these functional events occur. This tag is recorded to the genome, which triggers the creation of the relevant functionally mutated enzyme.

While the expansion of the gain-of-function biochemical topology is well justified as it represents the effects of any gain-of-function event within a genome or genome-like structure, the autocatalytic cycle of the paralogous gene is more difficult to rationalize. Here, I assume that a gene duplication event has occurred, and further posit that the duplicated gene is altered in a manner so as to maintain the original pathway's autocatalytic topology. Although such events are feasible, particularly if one assumes that this specific alteration to the duplicated gene is one that is highly probable upon the development and evolution of this biochemical pathway and is a specific functional mutation that is likely to occur across all enzymes, such assumptions significantly diminish the generalizability of this model. In addition, the alteration to the duplicated gene mirror the alterations made to the substrates as a result of the interactions from the normal, non-functionally mutated biochemical topology. Thus, the substrates of the A-#-R and the A-#-N enzymes retain structural similarities, with slight deviances.

Despite the clear drawbacks of using the extended topology with the paralogous genes,

using such a topology more importantly allows this thesis's aim to be completed, which is to examine the effects of an extended topology. The creation of new autocatalytic cycles, particularly in the manner used for this biochemical pathway, provides an interesting and simplified method of analysis for the evolution of complexity of biochemical pathways. To achieve a similar effect of allowing functionally mutated enzymes to create autocatalytic cycles, one would have to vastly increase the number of molecular species and interactions within the simulation, as entire autocatalytic cycles would be constructed using enzymes and substrates not initially related to the original biochemical pathway.

1.4 Black box approaches

For this project, although only the biochemical pathway of interest is explicitly modeled, the model implicitly accounts for the effects of various other molecules, particularly with regards to side-reactions and flux. Each substrate is given a probability, $P(r_{e\to i})$, that determines if a substrate is either used in a side-reaction or is diffused away from the proto-cellular environment. Using a random number generator, if the random number was greater than $P(r_{e\to i})$, the substrate is removed from the simulation. Similarly, each type of molecule was also given another probability, $P(r_{i\to e})$, which generated the said substrate and allowed it to appear in the simulation. This probability is responsible for modeling the diffusion of substrates into the proto-cell, as well as the creation of the said substrate due to a pathway not modeled within the simulation.

I selected the black box approach due to the impracticality of characterizing all

relevant molecules within a proto-cell. The interdependence of molecules within biological entities forces any attempt to view a particular biochemical pathway's effect on the cell to rely on multiple pathways that are not immediately of any interest to the experiment. The creation of an abstract yet valid biochemical network that defines all possible cellular interactions is limited by a) time constraints and b) lack of experimental data. Thus, the modeling approach on the effects of all non-explicitly (implicitly) modeled molecules in the simulation is a simplification grounded not only in technical feasibility, but also in the inability to a) successfully characterize abstract models to their real world counterparts and b) obtain sufficient data on the real world counterparts. Furthermore, the existence of these implicitly modeled molecules is already acknowledged through the usage of the Wiener process to model Brownian motion. Since we are only concerned with the existence and effects of the phenomena of side-reactions and flux on the explicitly modeled molecules, the black-box approach does not compromise the integrity of the model. It should be noted however that this particular approach then assumes that all other pathways and reactions within the model are operating *paribus ceteris*, that is, with little evolutionary change to their operations. This approach therefore accomplishes the objectives of this thesis, which is to examine the evolutionary changes within a proto-cell of a particular pathway, while compromising the model's ability to simulate possible real world proto-cells.

For the production of enzymatic agents from the genome, the black box approach was also used to avoid unnecessary complications within the model. While I realize that the molecular mechanisms of translation and transcription as well as the actual structure of the replicase and/or the replicase's existence have profound impacts on the model and

make certain assumptions on the type of proto-cell modeled, a simplified form of the genome was chosen as it was not the particular area of interest for this project. This enzyme production is functionally the same as gene expression, so long as the expression always leads to the production of the protein.

1.5 Performance values and performance mutations

This model uses an abstract quantitative concept of "performance" to quantify the proto-cell's ability to survive within a given environment as a result of the explicitly modeled pathway. The proto-cell's performance value increases or decreases based on the reactions that occur within the proto-cell. Certain reactions may result in products that are necessary in the functionality of the pathway's operations via their reactions with other molecules, and thus increase the pathway's contribution to the proto-cell. Similarly, other reactions may result in products that are deleterious to the operations of the pathway, and thus decreases the proto-cell's performance value. The contributions of reactions to the proto-cell's performance value are named "molecular performance values" (MPV) to differentiate itself from the proto-cell's performance value, which is the sum of all MPV and represent the functionality of the pathway's operations. It should be noted that it is not the existence of the molecule, but rather the creation or the destruction of particular molecules that affects these performance values. Since the MPV represents the state of functionality of membrane formation, and the modeled biochemical pathway is critical for membrane formation, it follows that a complete cycling of the biochemical pathway results in a net positive value, as the products of this pathway aid membrane formation.

The usage of a single numerical measure of the proto-cell's current state was

chosen as it provided a reliable method of tracking the effects of mutations on the proto-cell's state. The concept of performance values allows multiple independent pathways to be tested, as a weighted measurable output of the reactions is comparable across different reactions. While the abstract nature of the performance value is difficult to translate into the real world, it has the ability to create thresholds for determining the proto-cell's survival in the environment and ability to connect fundamentally different pathways with common molecules. This, I believe, is sufficient to justify its use.

Expanding upon this idea of performance values, the production of the enzymatic agents from the genome comes at a cost to the proto-cell's performance value. The genome is assumed to be constitutively expressing enzymes for simplifying purposes, as the effects of induced expression is not an objective of the paper. As a result, there are two competing forces that determine the proto-cell's performance value. The genome's constitutive production of enzymes results in a steady decline of the proto-cell's performance value, whereas the reactions from the biochemical pathway result in a positive net value. In addition to these forces, there is a third external factor that affects the proto-cell's performance value. To represent the effects of aging on the proto-cell, there is an additional factor that creates a steady decline of performance value. This decline in performance value, from both the aging factor and genome cost, dominates the net positive value from the biochemical pathway.

Under a pre-determined probability, the genome may create enzymes that have undergone a performance mutation, a mutation that affects the enzyme's overall performance and effectiveness to the membrane formation pathway. This is enabled within the simulation, as enzymes are hard-coded to add the MPV to the proto-cell's

performance value once they undergo a reaction. By allowing these hard-coded values of MPV to undergo performance mutation events, which create variance within the hard-coded MPVs that newly produced enzymes will contain, one can model the effects of real world mutations on the behaviors of enzymes. The effects of these mutations are based on a normal distribution with variance = 1 and the mean set to the MPV. In doing so, the initial MPV may undergo a performance mutation event, which alters the MPV to a new value. Further performance mutations then affect this new MPV, as the mean is set to the most recent MPV. Thus, cumulative mutations are enabled within this model.

To prevent these performance mutations from affecting already existing enzymes, two scripts are run in parallel to each other. The first script runs in the initial setup of the simulation and defines the characteristics of all of agents in the model prior to start of the simulation. The second script is exclusive to the genome and it is here that the performance mutator functions are contained. The second script operates only on the new enzymes that are being produced, allowing older enzymes to retain their MPVs.

To optimize the code and prevent needless loops, a performance mutator function exists for each and every enzymatic agent produced from the genome.

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;;; A-1-N Performance mutator ;;;
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to a-1-n-mutation-performance-1?
  if ((random-float 100) <= mut-prob-run)
  [set performance-1 random-normal a-1-n-performance-1 1]
end
```

Figure 6. Code for performance mutator function.

A normal distribution was selected to simulate the effects of real world mutations for practical reasons. Although a gamma distribution appears to more realistically portray the effects of real world mutations [24], the gamma distributions present in the ABM software used, NetLogo, do not permit cumulative mutations to occur.

1.6 Threshold concepts

To mimic the survival of the proto-cell, a survival threshold performance value (STPV) is used. If the performance value of the proto-cell is below the STPV, it suggests that the proto-cell's membrane formation capabilities are insufficient, and as a result, the increased concentration of molecular species without the expansion of the proto-cell will lead to the bursting of the proto-cell due to increased osmotic pressure. Since we are only interested in the characteristics of a functional proto-cell and the continuing of the simulation yields no practical data after death, the simulation stops once the performance value drops below the STPV.

If the proto-cell survives past a pre-determined time period since birth and the performance value exceeds another threshold value, the relative reproductive threshold (RRT), then the proto-cell will replicate. Here, performance values above this threshold suggest that the proto-cell's membrane formation machinery is operating optimally and thus able to allow normal expansion and replication to occur. The pre-determined time period since birth serves as a marker for defining when the proto-cell has reached sufficient maturity and thus expanded to the extent that self-replication due to osmotic

pressure will occur. Thus, the determining factor behind whether a proto-cell will self-replicate is whether the proto-cell's performance value is able to stay above the RRT for a select period of time. Since proto-cells are designed to have a decline in performance-value as time increases, the sharper this decline is, the less time the proto-cell will be able to spend above the RRT and the more difficult it is for the proto-cell to replicate. Alternatively, assuming that the decline in performance value is stable and approximately constant, the lower the RRT is, the proto-cell is able to spend more time above the RRT and can therefore reproduce much more easily. One can observe this relationship in Figure 6, where the intersection of the proto-cell's performance value under the high RRT occurs at roughly time-step 2, whereas the intersection of the proto-cell's performance under the low RRT occurs at time-step ~ 4.5. Thus, the larger the discrepancy between the initial performance value of the proto-cell and the RRT,ΔP_R, the more likely it is to survive. For this purpose, the initial performance value of the proto-cell was set at 500. To allow the progeny of the proto-cell to face the same ΔP_R, the performance value of the newly replicated proto-cell was set at 500 at birth, the same initial performance value of the parent proto-cell. The actual phenomenon of replication holds little value in this simulation, as a replication event does not change the operations of the proto-cell. Rather, what replication entails is a simple reset of the proto-cell's performance value to 500, an increased probability of mutation for a set time period after replication, and a marking of the replication event in the simulation.

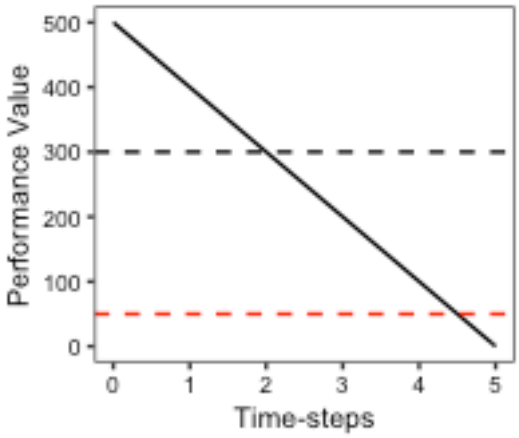


Figure 7. Effect of RRT on proto-cell survival. Dashed lines indicate hypothetical RRT while the solid line represents a hypothetical proto-cell's performance. Black (high) RRT is at performance value = 300 while the red (low) RRT is at performance value = 50.

The idea of a threshold is contingent on the validity and interpretation of the performance value earlier described. Since the performance value is a relative measure of the molecule's utility to the cell's ability to survive in a given environment, the threshold concept must similarly be a relative measure that describes the difficulty or ease of the proto-cell to survive or replicate in its given environment. Thus, these threshold values will change based on the MPV provided by each reaction type as well as the desired level of difficulty for proto-cell survival. As a result, threshold values allow the researcher to define the environmental context of the proto-cell in terms of performance value. Threshold values also play a powerful role in determining how the parameter space is explored. Premature deaths limit the number of possible mutations and therefore prevent the proto-cell from exploring the parameter space while unnecessary longevity of survival results in weak selection pressure. In addition, the halting of the simulation based on proto-cell "death" is analogous to the destruction of the proto-cell due to its inability to

stably maintain itself within a given environment and therefore represent end-points for self- replicating biological entities.

1.7 Molecular movement

To model molecular movement, I used Brownian Dynamics integrated with a discrete time Euler-Maruyama scheme as a random walk, resulting in an expected mean squared displacement of:

$$x_j(t + \Delta t) = x_j(t) + \sqrt{2D_j \Delta t}\, \xi \tag{1.0}$$

where ξ is a 2-dimensional zero-mean Gaussian random variable with unit variance and D_j is the diffusion coefficient [6, 25]. The diffusion coefficient may also be modeled via the Stokes-Einstein equation:

$$D_j = \frac{k_B T}{6\pi\eta r_j} \tag{1.1}$$

where k_B is the Boltzmann constant, T is temperature, η is the viscosity of the cytoplasm, and r_j is the hydrodynamic radius of the molecule.

Several assumptions are made in modeling molecular movement. Despite their importance in real-world molecular movement, intramolecular forces and potentials are ignored. Furthermore, complicating factors such as crowding factors within the cell (from the existence of particular cellular structures) are also abstracted away for simplifying purposes.

1.8 Enzyme kinetics

To simulate a biochemical pathway within a proto-cell, it is necessary to model the proto-cell's intracellular kinetics. I have chosen a stochastic approach, using a reaction probability (probability that the collision of two molecules yields a reaction) and binding time (time between formation of enzyme-substrate complex to dissociation) to model reaction kinetics as it may be valid for approximating real-world reactions [26]. In doing so, parameters such as steric factors were abstracted while maintaining experimental integrity.

For a collision to occur, the molecule, defined in size by its hydrodynamic radius, must enter the space of another molecule's hydrodynamic radius. Once a collision has occurred, a random number from 1 to 100 was generated. If this random number falls below the reactive collision probability parameter, then the collision will initiate a reaction.

To do so, velocity for a single molecule is converted into speed and direction (heading representations) components along and perpendicular to an angle theta.

$$v_x = speed * \cos(\theta - direction) \quad (2.0)$$

$$v_y = speed * \sin(\theta - direction) \quad (2.1)$$

When the collision occurs, the velocity of the center of mass (v_{cm}) along theta will be calculated and the new speed and heading representations are calculated by

$$v_x' = 2 * v_{cm} - v_x \quad (2.2)$$

$$v_y' = 2 * v_{cm} - v_y \quad (2.3)$$

The above velocities will then be converted back into a speed and direction for the molecule to travel at. To do so, the speed will be calculated via:

$$speed = \sqrt{v_x^2 + v_y^2} \quad (2.4)$$

and direction is represented via:

$$direction = \theta - \tan^{-1}(\frac{v_y}{v_x}) \quad (2.5)$$

If the two colliding molecules are an enzyme and substrate, a random number from 1 to 100 will be generated. If this random number falls below the reactive collision probability parameter, then the collision will initiate a reaction. Similarly, if the number is above the parameter, a reaction will not occur and the molecules will bounce off each other, with their speed and direction affected by a change in kinetic energy of the molecule.

After the simulation decides that a reaction has indeed occurred, an enzyme-substrate complex is formed. To do so, the substrate disappears from the simulation while the enzyme molecule is transformed into a complex, with an added mass of both the enzyme and substrate to identify the creation of a successful complex. Upon the creation of this complex, the simulation records the current time-step. In addition, the simulation marks the complex with a complex-identifying tag to prevent the complex from reacting with other substrates and to allow the following complex-related procedures to occur. The simulation then subtracts the time-step of the simulation with the recorded time-step (from the time of complex formation). Once the difference between the current time-step and

recorded time-step is greater than a pre-determined binding complex time parameter, the complex dissociates and is converted back into the enzyme agent. The dissociation of the complex then triggers the creation of a new product agent within the simulation.

1.9 Molecule characteristics

The characteristics of each enzyme were identical except for the performance value and the substrate and product. Although heterogeneous enzymatic interactions follow different dynamics and can impact the behavior of autocatalytic systems, this vastly simplifying assumption was used to reduce the subset of possible parameters to manipulate within the model. For all runs of the model, there were 30 of each type of enzyme and 300 of each type of substrate within the initial setup of the model. All enzymes faced a 60% probability of yielding a successful reaction with a binding-time of 15 time-steps. All molecules also had a radius of .3 patches and had a mass of value 1. Furthermore, all runs of the model, excluding the topology- and mutation-related runs were unable to undergo both functional and performance mutations unless otherwise stated. Under all runs, the time requirement for replication was 40 time-steps above the RRT. The genome and aging cost was -4 performance value per enzyme created, with 1 of each enzyme produced every 3 time-steps for a total of -20 performance value per 3 time-steps. The autocatalytic cycle produced a net positive of +28 to the performance value upon one full cycling of the biochemical pathway.

Chapter 2

Experimental Results

To estimate the parameters for a working proto-cell, an additive one-factor analysis was used. Once this single parameter is optimized according to minimal conditions of the model, I selected another parameter to optimize and so on, until all parameters are optimized for a single local optima.

2.1 Validation of modeled enzyme kinetics

To establish the validity of characterizing enzyme kinetics through reaction probabilities and binding time, a Michaelis-Menten plot of the enzyme's activity was generated.

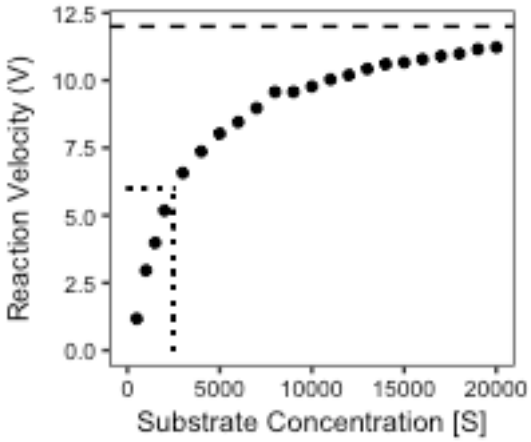


Figure 8. Michaelis-Menten plot of the A-1-N enzyme. Simulation occurred with a fixed number of 30 A-1-N enzymes with varying concentrations of substrate$_1$. V_{MAX} is roughly 12.5 products per time-step and K_M is roughly 2500 substrates. Reaction probability upon collision was set at 60% and binding-time was set at 15 time-steps.

Here, one can observe the similarities between the simulated enzymatic behavior and real world enzymes. Due to computational costs, an upper limit of 20,000 substrates within the simulation was created. The steep slope within low concentrations of the substrate suggest noticeable gains in reaction velocity, or the amount of products a single enzyme may produce on average in a single time-step. However, as substrate concentrations increase, there is a diminishing marginal rate of reaction velocity due to a decreasing number of 'free' A-1-N enzymes and a higher number of A-1-N and substrate$_1$ complexes as a result of lower average search times for enzymes to find a substrate with greater substrate concentrations [27]. Since the enzymes within the simulation only differ in what they react to, instead of having more fundamental differences, the enzymatic activity displayed in Figure 8 is applicable to all enzymes within the model.

2.2 STPV and RRT

To select a STPV for the model, I tested the STPV value of -10,000 under 100 runs. Since the STPV determines the survival of the proto-cell, and falling below the STPV merely ends the simulation compared to altering the proto-cell's operations or behavior, it was not necessary to test the STPV under various intervals. Rather, by selecting a large STPV value and observing the behavior of the proto-cell at various points of its life, one can then determine if a particular STPV value would yield the desired environmental stress upon the proto-cell. To see if the STPV value would be appropriate for the model, I measured the number of time-steps that the proto-cells were able to survive within the environment, with an upper bound of 5,000 time-steps.

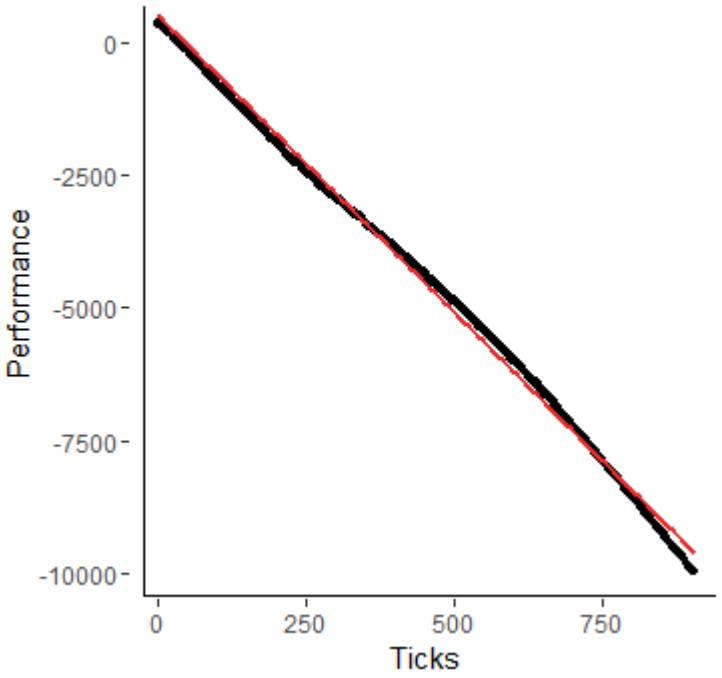


Figure 9. Survival of proto-cell. STPV is set at -10,000. Simulation ended at ~ 900 ticks. Red line represents linear regression on the model, whereas the black line represents the average performance values of 100 proto-cells.

From Figure 9, one can observe the low variability in performance values due to the degree of variance of the performance values from the linear regression. I decided to choose a STPV of -2000 as it provided an environment that allowed the proto-cell to survive without compromising on selection pressure, as such proto-cells would tend to survive for approximately 80 time-steps, a time-horizon that works well for the scope of this project.

For the RRT selection of the ideal RRT was based upon the number of self-replicating generations that could form. I tested the RRT from 400 to -500 in intervals of

100. To reduce degree of variability within the experiment, each RRT value was tested through 100 runs.

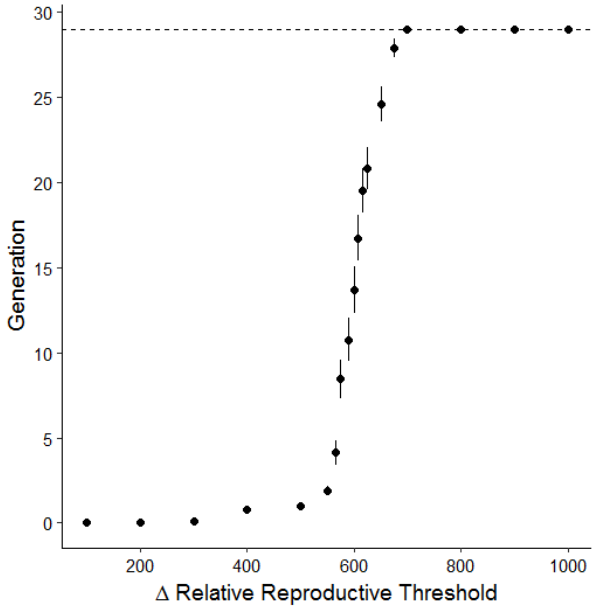
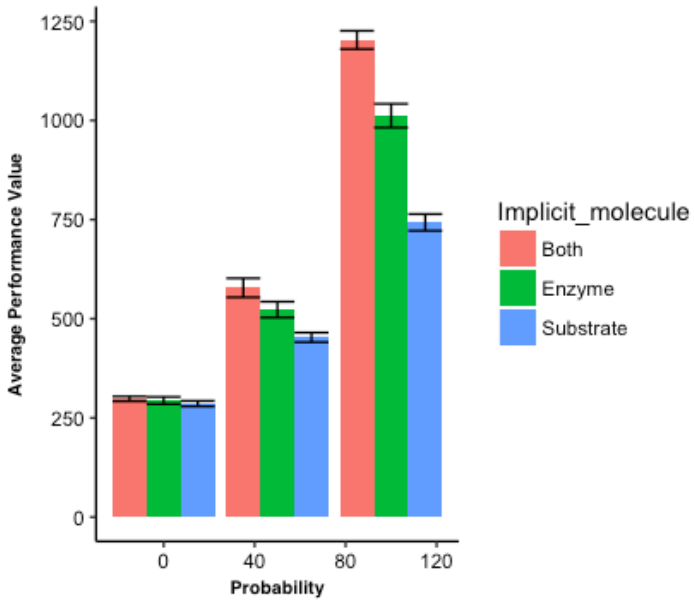


Figure 10. Reproductive ability of proto-cell under environmental stress. Relative reproductive threshold values indicate proto-cell's ability to reproduce itself. Generation values are averages of 100 runs. Maximum number of ticks per simulation run = 1200, with a maximum possible number of 29 generations (indicated by dashed line). Values are mean ± SEM.

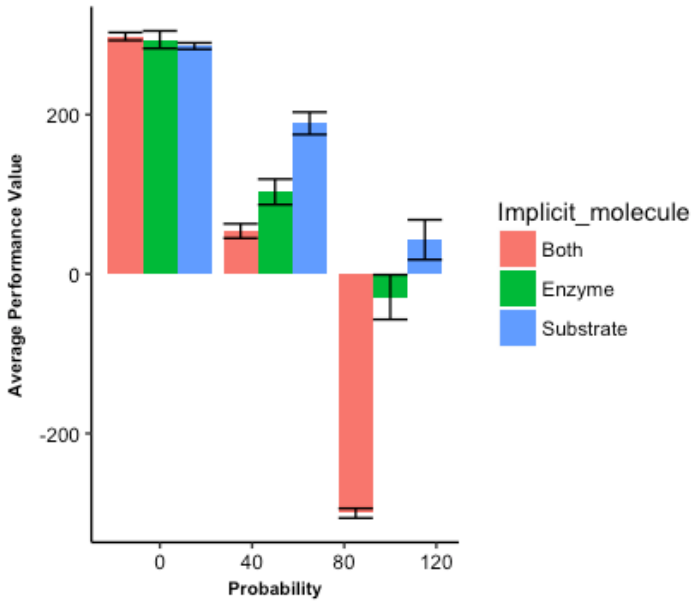
2.3 Side-reactions, flux, and rate of enzyme production

For the purposes of this model, side-reactions, flux, and enzyme production from the genome are functionally equivalent. Each of these phenomenon results in the appearance of molecules within and/or out of the proto-cell. The sole difference between side-reactions and flux versus enzyme production is the genome cost associated with enzyme production.

To choose an appropriate probability of $P(r_{e\to i})$ and $P(r_{i\to e})$ for the substrate, we analyzed the effects of the probabilities on performance value. To test the effects of $P(r_{e\to i})$ and $P(r_{i\to e})$, a one-factor analysis was taken. There were 3 sets of experiments, each accounting for the appearance or disappearance of particular subsets of molecules and testing the probabilities of 0%, 50%, and 100%. The first set of experiments dealt with the appearance and disappearance of enzyme molecules. The second set of experiments dealt with the appearance and disappearance of substrate molecules. The third set of experiments dealt with the appearance and disappearance of all molecules (excluding the genome). Each experiment underwent 10 trials. As expected, higher values of $P(r_{i\to e})$ for substrates, enzymes, and both, resulted in higher performance values and conversely, higher values of $P(r_{i\to e})$ for substrates, enzymes, and both lead to lower performance values.

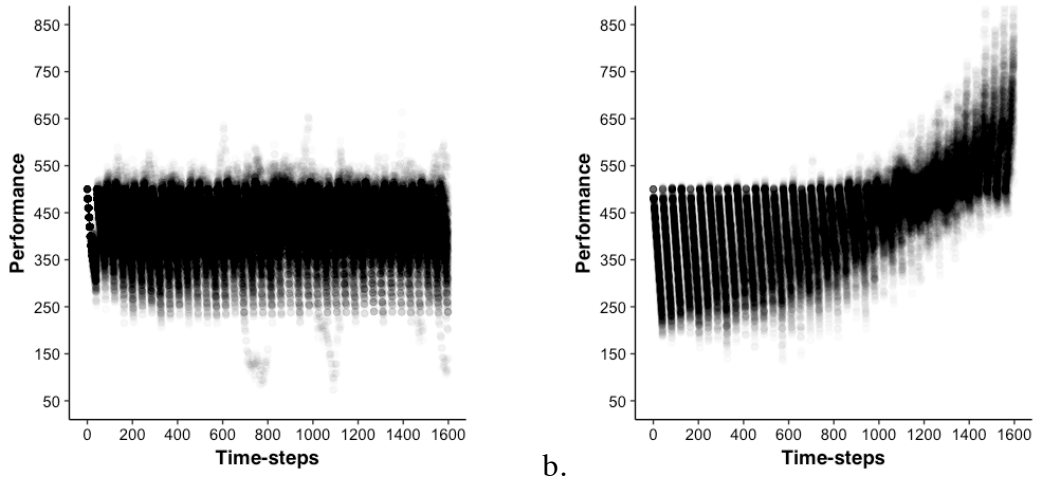


a.



b.

Figure 11. Effects of flux/side-reactions on performance value. Figure 11a. shows the effects of flux when the probabilities of molecules appearing within the simulation are manipulated. Probabilities are 0%, 50%, and 100%. Color indicates type of molecule whose probability of appearing is manipulated. Error bars represent mean ± SEM. Figure 11b. shows the effects of flux when the probabilities of molecules removed from the simulation are changed. Probabilities are also 0%, 50%, and 100%.



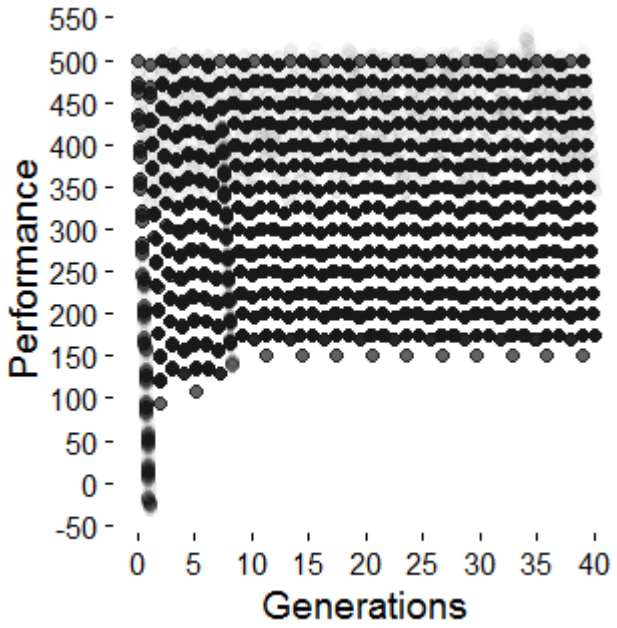
a. b.

Figure 12. Non-linear effects of gene expression on performance. Figure 12a shows the performance of 100 proto-cells capable of undergoing functional and performance mutations. The genome produces five of all the enzymes per 3 time-steps. Figure 12b shows the performance of 100 proto-cells also capable of functional and performance mutations with a higher rate of gene expression (5 enzymes per time-step) for all enzymes. Gene expression is the rate of production for enzymes. Darker areas indicate greater density (similar performance values across different proto-cells at the same time). Relative reproductive threshold is less than 0. Self-replication occurs at 40 time-steps if performance > RRT. RRT < 0. Each circle represents a proto-cell's performance value at a particular time-step.

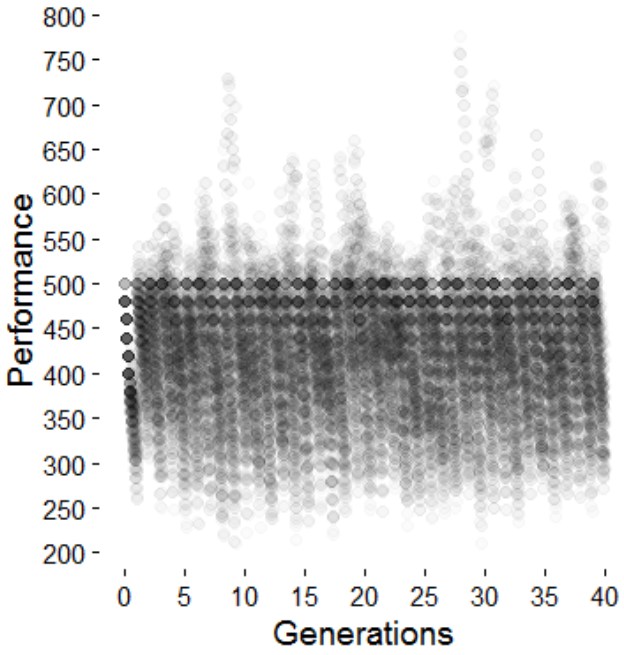
At higher probabilities of $P(r_{i\to e})$ and higher rates of enzyme production, an interesting phenomenon occurred. As seen in Figure 12, higher rates of enzyme production lead to a nonlinear increase in performance values.

To differentiate between crowding effects (increased number of enzymes relative to substrates) and the effects of the autocatalytic nature of the pathway, I repeated the

experiment but did not allow A-5-N to create $substrate_1$ as a product. Rather, A-5-N created a unique substrate, $substrate_6$, which eliminated the autocatalytic nature of the pathway but with a larger than normal amount of enzymes. Here, the benefits of an autocatalytic cycle are clear.



a.



b.

Figure 13. Increased gene expression in a non-auto-catalytic pathway. 13a. shows the performance of 100 proto-cells capable of undergoing performance, but no functional mutations. For these proto-cells, A-5-N does not create substrate$_1$ as a product, and thus makes the pathway non-auto-catalytic. Enzyme production occurs at 5 enzymes per 3 time-steps. 13b. shows the same conditions as 14a but has increased enzyme production (5

enzymes per time-step). Generations are 40 time-steps, created here to more easily identify when a generation occurs. RRT < 0. Each circle represents a proto-cell's performance value at a particular time-step.

One can observe that although there is indeed an increase in the average performance value of the proto-cell due to the increase of enzymes, there exists no non-linear trend as seen in Figure 12. Furthermore, it is interesting to note the lack of variance that exists within Figure 13a., where proto-cells appear to follow near identical paths. This abnormal characterization of proto-cell behavior across independent trials suggests either a systematic flaw within data collection or that low levels of enzyme production/low quantities of enzymes and a non-auto-catalytic pathway limit the variances in behavior amongst proto-cells.

2.4 Probability and strength of mutations

For this series of experiments, I increased both the probability and strength of mutation per experiment. Probabilities ranged from 0 to 100% chance of mutation per time-step, with experiments running in intervals of 50%. In parallel to these probabilities, the strength of the mutations, represented by the variance parameter of the normal distribution of the performance mutator function, ranged from 0–4, with an increase of 2 per experiment. For these experiments, 20 trials were run for each condition.

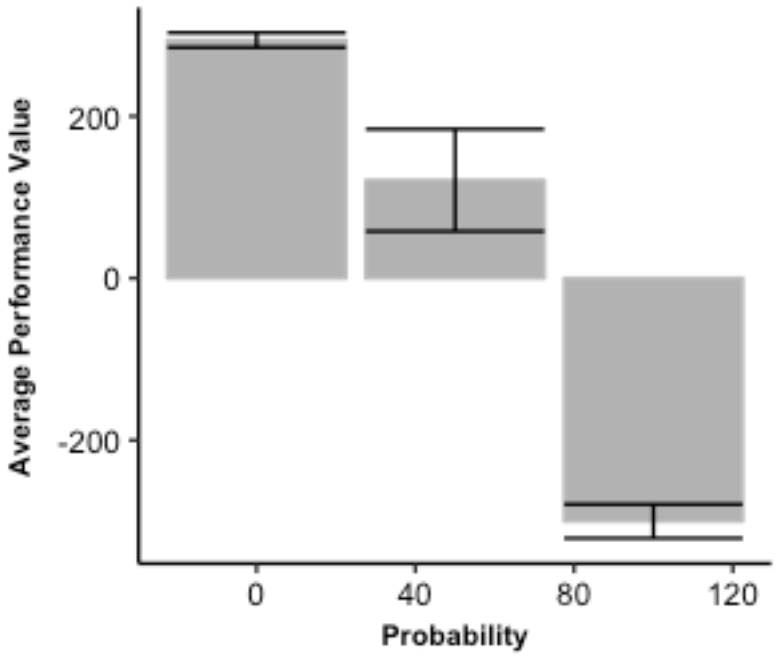


Figure 14. Average performance values under varied performance mutations. Probabilities of a performance mutation occurring are 0%, 50%, and 100%. The strength of the mutation depends on a normal distribution, and the variances for each group of performance mutation probabilities are 0, 1, and 2 respectively. Error bars represent mean ± SEM.

Moderate likelihood and strength of mutations create the greatest variance, whereas the groups with the most and least likelihood and strength of mutations had the least variance. In terms of average performance value, there appears to be an inverse relationship of with greater likelihood and strength of mutation. However, it should be noted that within the 100% mutation group, a large majority of them died, drastically distorting the average performance value and minimizing variance.

2.5 Topology and Complexity

To understand the effects of biochemical topology on the development of the proto-cell, I compared the original pathway to the pathway with the gain-of-function mutation and the paralogous gene functional mutation.

Similar to the results found in the increased rates of enzyme production in Figure 12, Figure 15 shows a non-linear trend due to the increased autocatalytic nature of the expanded biochemical topology.

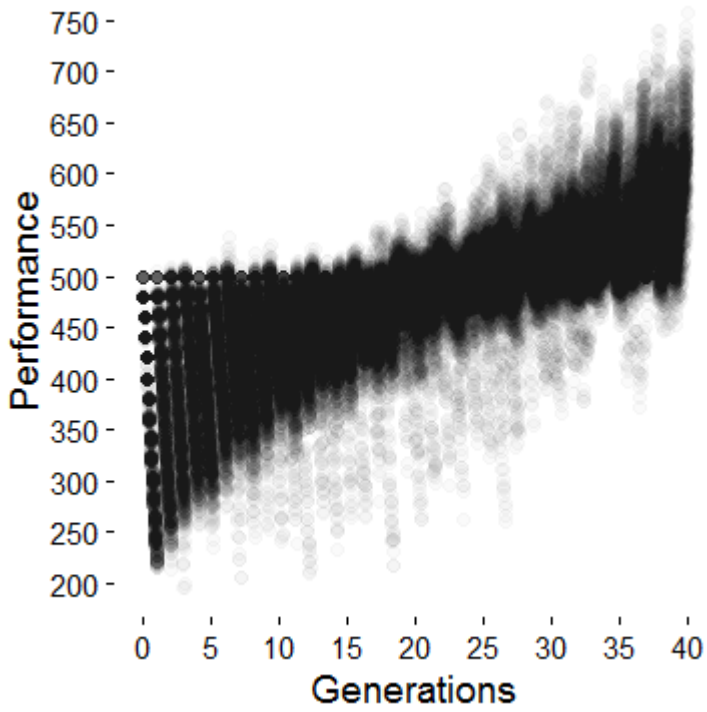
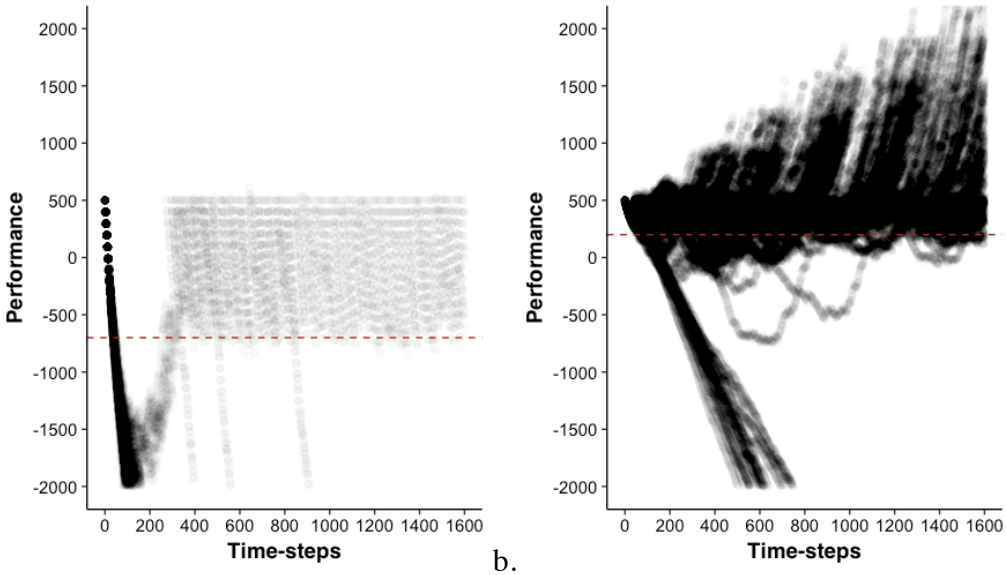


Figure 15. Typical proto-cell profile under the expanded biochemical topology. The probability of undergoing functional mutation events is 20% per time-step. RRT is below

0. Self-replication occurs after 40 time-steps if the performance value > RRT. Each circle represents a proto-cell's performance value at a particular time-step.

With particular conditions, I was able to identify cases where proto-cells were able to recover from otherwise fatal environments due to their ability to undergo functional mutations.



a. b.

Figure 16. Recovery of proto-cell self-replication ability due to functional mutations. Dashed line indicates relative reproductive threshold. Proto-cells are able to fully evolve in 16a and 16b. In figure 16a, evolution probability is high. In figure 16b, probability of mutations and evolution are low. Number of enzymes is also limited in figure 16b to what initial parent cell had (to avoid non-linear effects) and more realistically model proto-cells. Replication occurs if performance value > RRT after 40 time-steps. Each circle represents a proto-cell's performance value at a particular time-step.

When comparing the averages of the original pathway, an extended form of the pathway due to the gain-of-function mutation, and the final pathway with both the gain-of-function and paralogous gene functional mutations, the following relationship is seen in Figure 17.

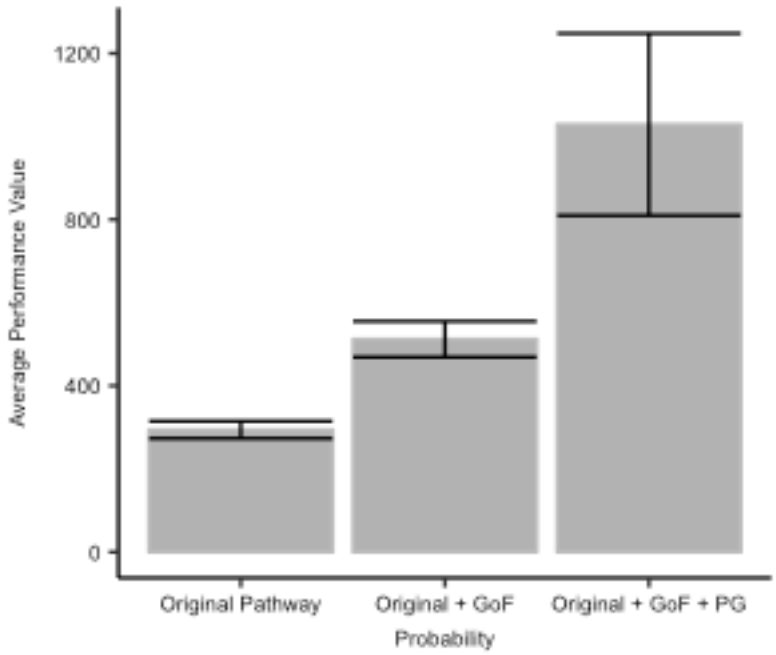


Figure 17. Averages performance value per biochemical topology. Averages are of 100 runs. Original pathway is defined as the pathway in black in Figure 5 and has no functional mutations. Original + GoF represents the original pathway with the gain-of-function mutation enabled for all enzymes. Original + GoF + PG represents the original pathway with both the gain-of-function and the paralogous gene functional mutations. Error bars represent mean $\pm$ SEM.

Chapter 3

3.1 Discussion

In this work, I developed an agent-based mesoscopic model of an abstract proto-cell. This model demonstrated the benefits of an extension of a biochemical topology in a proto-cell. Similarly, the model also shows the effects of flux on particular subsets of molecules and its implications on the formation of early life. The model also has implications on some of the necessary environmental conditions for the formation of stably self-replicating proto-cells.

Interestingly, according to Figure 10 there exists a non-linear relationship between the differential, ΔP_R, of the RRT and initial performance value versus the number of generations that the proto-cell could self-replicate. If ΔP_R is sufficiently large, the proto-cell is able to stably replicate for the entirety of the simulation. However, if ΔP_R is sufficiently small, then the proto-cell is either unable to maintain the necessary performance value for reproduction or will die in a few generations due to the accumulation of deleterious mutations. Furthermore, there exists a narrow window where

slight changes to RRT determine if the proto-cell is able to stably reproduce itself *ad infinitum* or if the proto-cell's progeny will not reproduce, despite no changes in the environmental stress. However, outside of this window, changes in RRT appear to have little to negligible effects on the number of generations the proto-cell is capable of producing. Therefore, relatively insignificant changes to the proto-cell's machinery (which affect the rate of performance value decline and thus the time the proto-cell spends above the RRT) or minor changes in the environment may lead to drastic changes within the proto-cell's ability to be stably self-replicating *ad infinitum* (in the absence of any unexpected events).

According to Figure 16 and Figure 17, recovery of the proto-cell through functional mutations is possible. Furthermore, the non-linear effect of increased enzyme production in autocatalytic cycles suggests that either small changes with the environmental flux of molecules or side-reactions (as a result of the phenomenological similarities with enzyme production) or rate of gene expression may have major impacts on the development of the proto-cell. The autocatalytic nature of the original pathway and the addition of a new autocatalytic pathway appear to have profound effects on the performance values and thus the efficiency of the membrane formation machinery for the proto-cell.

In addition, according to Figure 11, flux/side-reactions appear to have the heaviest impact when both enzymes and substrates are affected. However, it should be noted that the removal of enzymes from the simulation appears to have a larger effect than that of the removal of substrates, likely due to the autocatalytic nature of the pathway. As expected,

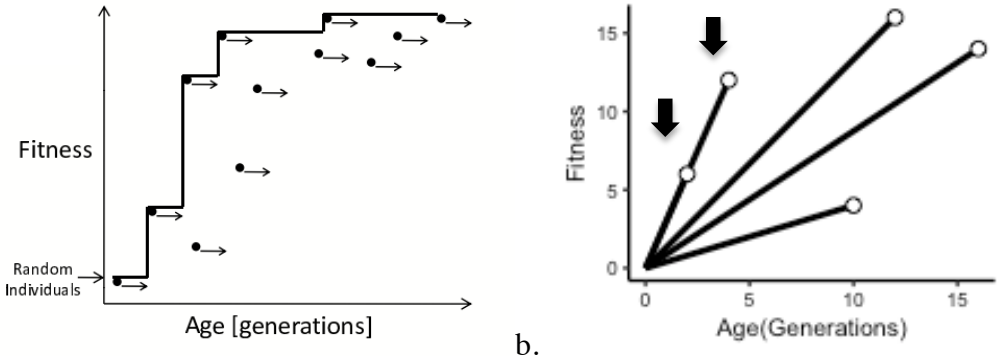
the removal of enzymes and substrates had a negative impact on the performance of the proto-cell and vice versa.

Greater mutation occurrence and strength also appear to have a negative effect on the development of proto-cells, although it should be noted that very wide intervals were taken in this particular project. As a result, there may have been optima for mutations that were missed. At least within the extremes of great mutation occurrence and strength, the variability killed off a majority of proto-cells.

There exist several limitations with the experimental design of the model. The author is aware of the abstract nature of the performance value and the difficulty in procuring a real-world equivalent single measure that is capable of describing a biological entity's ability to function in an environment. The particular method used to model performance values also lacked time-horizons where products may, through downstream interactions, benefit the proto-cell through a series of side-reactions. Furthermore, it is likely that products will undergo varied side-reactions, where different probabilities dictate the occurrence of these side-reactions. Thus, with the existing experimental design, it is not possible to model the effects of various side-reactions, as they are implicitly modeled as a single overarching side-reaction under the current black-box approach. In addition, the vast parameter space and lack of experimental data on a mesoscopic scale greatly increased the labor-intensity and time necessary for parameter estimation. Parameter estimation was also limited to either a one- or two-factor analyses, which may be wholly insufficient for related or causally linked parameters.

To avoid these issues, the AFPO algorithm should be used in the future. The AFPO algorithm is used for *post-hoc* analysis to identify local optima and/or identify

conditions that allowed the proto-cell to stably survive given a particular SPTV. This AFPO algorithm selects for individuals with greater fitness (performance value in this instance) and lower age (generations and time-steps since birth), as competitive younger individuals are able to maintain greater diversity in further generations relative to older individuals with similar performance values. In contrast to the AFPO approach, the additive one-factor analysis is a more elementary method where a single parameter is tweaked until local optima is identified.



a. b.

Figure 18. Optimization via the AFPO algorithm. In Figure 18a., optimal individuals are those amongst the solid line, which are the youngest in the set of individuals while retaining the highest fitness. Arrows indicate the possible further generations that the individual may produce. Once a subset of optimal individuals are identified, to identify the most optimal individual, one must look at which individual has the greatest fitness per age. Figure 18b shows an optimal selection of individuals, with arrows showing the most optimal individuals. Figure 18a. is from Schmidt and Lipson, 2010.

Since the AFPO algorithm is *post-hoc* and there exists a large parameter space due to the number of parameters employed, I would select random sets of parameters to test.

Once I identified a working set of parameters, I selected a particular parameter and varied its value to a high and low extreme, to observe its effect on the model. An optimal collection of parameters would create a stable performance value, where the stochastic noise of the model is minimized and where mutations are limited but have a profound influence on the proto-cell. Although I had originally intended on using the AFPO algorithm to more deeply investigate the particular parameters and combination of parameters that lead to a proto-cell's survival, time constraints meant that a full usage of the AFPO algorithm could not be used, and incomplete data from the AFPO algorithm meant that a thorough understanding of the parameter sets could not be achieved. Thus, the AFPO algorithm is not utilized in this thesis.

However, development of some of the black-box approaches used in this model, particularly with regards to the isolation of a single biochemical pathway of interest, may prove useful to future research. Further expansion of this model, or experimental validation of it, may aid in a mechanistic understanding of how particular pathways have evolved, and the modeling of biochemical pathways necessary for the proto-cell or for the origins of life may yield interesting results. For example, compartmentalization of the cell, a more comprehensive modeling of early replicases, or environmental effects on the proto-cell may expand our understanding of why certain structure or pathways have formed and why other theoretical structures or pathways fail to be realized in the real world.

References

1. Coveney PV, Fowler PW. Modelling Biological Complexity: A Physical Scientist's Perspective. J R Soc Interface. 2005; 2 (67): 267-80.

2. Wilensky U, Rand W. An Introduction to Agent-based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. Cambridge: MIT Press. 2015.

3. Bonabeau E. Agent-based Modeling: Methods and Techniques for Simulating Human Systems. Proc Natl acad Sci U S A Supplement. 2002; 3: 7280-287.

4. Tong X, Chen J, Miao H, Li T, Zhang L. Development of an Agent-Based Model (ABM) to Simulate the Immune System and Integration of a Regression Method to Estimate the Key ABM Parameters by Fitting the Experimental Data. PLoS One. 2015; 11 (5): n. pag.

5. Rahmandad H, Sterman J. Heterogeneity and Network Structure in the Dynamics of Diffusion: Comparing Agent-Based and Differential Equation Models. Manage Sci. 2008; 54 (5): 998-1014.

6. Klann M, Koeppl H. Spatial Simulations in Systems Biology: From Molecules to Cells. Int. J. Mol. Sci. 2012; 13 (12): 7798-827.

7. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolivar B, Assad-Garcia N, Glass JI, Covert MW. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell. 2012; 150: 389-401.

8. Wolfram S. A New Kind of Science. Champaign: Wolfram media; 2002.

9. Feig M, Sugita Y. Reaching New Levels of Realism in Modeling Biological

Macromolecules in Cellular Environments. J Mol Graph Model. 2013; 45: 144-56.

10. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC. Essential genes of a minimal bacterium. Proc Natl acad Sci U S A. 2006; 103(2): 425-430.

11. Schrodinger E. What is Life?: The Physical Aspect of the Living Cell. Cambridge: Cambridge University Press; 1944.

12. Ganti T. The principles of life. With a commentary by J. Griesemer and E. Szathmary. Oxford: Oxford University Press; 2003.

13. Bahadur K. Synthesis of Jeewanu, the Protocell. Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene. Zweite naturwissenschaftliche Abt.: Allgemeine, landwirtschaftliche und technische Mikrobiologie. 1966; 121(3):291-319.

14. Eigen M, Schuster P. The hypercycle: a principle of natural self-organization. Springer Science & Business Media; 2012.

15. Railsback SF, Grimm V. Agent-based and individual-based modeling: a practical introduction. Princeton: Princeton University Press; 2011.

16. McGraw Hill Encyclopedia of Science and Technology. 8th ed. New York: McGraw Hill;1997.

17. Bicknese S, Periasamy N, Shohet SB, Verkman AS. Cytoplasmic Viscosity Near the Cell Plasma Membrane: Measurement by Evanescent Field Frequency-Domain Microfluorimetry. Biophys J. 1993; 65: 1272-128

18. Hordjik W, Hein J, Steel M. Autocatalytic Sets and the Origin of Life. Entropy. 2010; 12: 1733-1742.

19. Vasas V, Fernando C, Santos M, Kauffman S, Szathmáry E. Evolution before

Genes. Biology Direct. 2012; 7 (1): 1-14.

20. Ganti T. On the early evolutionary origin of biological periodicity. Cell Biology International 2002; 26 (8): 729-735.

21. Pereto J. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. Chem. Soc. Rev. 2012; 41: 5394-5403.

22. Dearmer D, Dworkin JP, Sandford SA, Bernstein MP, Allamandola LJ. The first cell membranes. Astrobiology. 2002; 2: 371-381.

23. Orgel LE. Prebiotic Chemistry and the Origin of the RNA World. Critical Reviews in Biochemistry and Molecular Biology. 2004; 39: 99-123.

24. Eyre-Walker A. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics. 2006; 173 (2): 891-900.

25. Ermak DL, Mccammon JA. Brownian Dynamics with Hydrodynamic Interactions. J Chem Phys. 1978; 69 (4): 1352-380.

26. Pérez-Rodríguez G, Pérez-Pérez M, Glez-Peña D, Fdez-Riverola F, Azvedo NF, Lourenço A. Agent-Based Spatiotemporal Simulation of Biomolecular Systems within the Open Source MASON Framework. Biomed Res Int. 2015: n. pag.

27. Tong D. Kinetic Theory: University of Cambridge Graduate Course. Unviersity of Cambridge; 2012

28. Schmidt M, Lipson H. Age-Fitness Pareto Optimization. In: Riolo R, McConaghy T, Vladislavleva E., editors. Genetic Programming Theory and Computation VIII. Vol:8. New York: Springer; 2010. pp. 129-146

Supporting Information

Code_S1.pdf