A TOY MODEL FOR THE EVOLUTION OF DNA POLYMERASE POLARITY

by

Joshua Ballanco

A DISSERTATION

Submitted to the Faculty of the Stevens Institute of Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

_____

Joshua Ballanco, Candidate

ADVISORY COMMITTEE

_____

Marc Mansfield, Chairman   Date

_____

A. K. Ganguly   Date

_____

Knut Stamnes   Date

_____

Nicholas Murgolo   Date

STEVENS INSTITUTE OF TECHNOLOGY
Castle Point on Hudson
Hoboken, NJ 07030
2009

A TOY MODEL FOR THE EVOLUTION OF DNA POLYMERASE POLARITY

ABSTRACT

It is my thesis that life can be modeled as a thermodynamic system. Specifically, it is my goal to use this approach to understanding life in order to investigate evolution as an energy minimization problem. In this sense, ecological niches represent local energy minima, selection criteria are nothing more than potential energy gradients, and extinction is the rate of change of the energy landscape exceeding the kinetics of adaptation. This idea is derived from the observation that life is, at its most fundamental level, nothing more than an elaborate collection of chemical reactions. Taken individually, each of these reactions is governed by the laws of thermodynamics, and thus far there is no known upper limit to the scale to which thermodynamic principles may be applied.

In this work I present a model systems which is inspired by this hypothesis. In this model, I will look at the directionality of nucleotide polymerases, all of which synthesize new nucleotide polymers in a 5' to 3' direction. This phenomenon could be the consequence of a very early founder effect. On the other hand, it could be that this directionality evolved due to an inherent advantage. I would propose that the process of nucleic acid polymerization provides a clue as to what that advantage might have been. The toy model I have constructed demonstrates this, and the results provide an insight into the forces that might have driven early biological evolution.

Author: Joshua Ballanco

Advisor: Marc Mansfield

Date: Sept. 18, 2009

Department: Chemistry, Chemical Biology, and Biomedical Engineering

Degree: Doctor of Philosophy

This thesis is dedicated to my loving and supportive wife.

Acknowledgements

I would like to acknowledge Marc Mansfield for the years of guidance, inspiration, and instruction that he so graciously provided during both my undergraduate and graduate education. I would also like to acknowledge Francis Jones for making this research possible, for the encouragement that he provided me as an undergraduate and the opportunity he made available to me as a graduate. I would also like to thank the members of my committee, A. K. Ganguly, Knut Stamnes, and Nicholas Murgolo, for their input and invaluable insight. Thanks also go out to the faculty and staff of the Department of Chemistry, Chemical Biology, and Biomedical Engineering for all the help they have provided through the years, and most especially to Mary Newell without whom I would not have made it past the first day of graduate school. Finally, I must humbly admit that the thesis presented here would not exist if it were not for the constant support and encouragement of my beautiful wife Ceren Örnek-Ballanco.

Table of Contents

List of Figures

Chapter 1

Introduction

Evolution has been, since the time of Darwin, a science mostly concerned with the past. The principles of evolution have been used to explain observations of ancient creatures and to predict what new evidence from the past should eventually turn up. One example of this sort of prediction is the existence of transitional forms. Evolution through the gradual acquisition of altered traits would predict the existence of ancient animals with a blending of traits from closely related branches in the tree of life. Indeed many such forms have been discovered, validating this prediction.

When it comes to predicting the past, evolution has been wildly successful. Where the study of Evolution has been thus far lacking is in its ability to predict the future. That is, looking at collected fossils and current natural evidence, the Theory of Evolution gives us the ability to identify which selective pressures acted on past populations. What the Theory of Evolution cannot do, at present, is predict which environmental or other influences will act as selective pressures going forward. At best, we can make educated guesses based on past evidence, but we lack even the basic ability to assign a concrete measure of confidence in such predictions.

This situation is akin to a meteorologist being able to explain why it rained yesterday but completely unable to predict tomorrow's weather. It is important, though, to understand that this is not an inherent failing of evolution, but rather a sign of a young science with much work left to do and many discoveries yet to be made. In physics, Newton's laws were successful at explaining the action of an apple falling from a tree for many centuries before they were developed to the point they could inform us how to send a man to the moon. What's remarkable about this development,

from falling apples to moon landings, is that it did not require any fundamental additions to Newton's laws. (Certainly General Relativity has fundamentally altered Newton's laws, but it is not, strictly speaking, an alteration required to get to the moon.) Rather, physics was able to make this progress merely through improved tools, improved instruments, and most importantly, improved means of applying the fundamental laws to a problem.

Thus, in order to advance the field of evolution, we should strive for more and better data, but also new and better ways of analyzing and testing that data. Before looking at how this might be achieved, let us look at what the implications of better predictive power in evolution might be. A rather straight forward implication would be the ability to predict the occurrence and course of epidemic or even pandemic diseases. Such diseases are biological organisms subject to Darwinian evolution and, often times, humans are the niche for which they are adapting. We very frequently alter their niche, introducing new selective pressures, in the ways that we treat disease with medicines, quarantine, or myriad other techniques.

The implications can be more far reaching then they might first seem. Partially, this is due to the fact that evolution, and the principles of Darwinian Evolution, apply to a much more diverse range of situations than just the origin of animal and plant species. For example, cancer is an evolutionary process. With each round of chemotherapy or radiation treatment, those cancer cells that have adaptations that increase their resistance to treatment will be more likely to survive. These cells will, thus, seed what will almost inevitably form as a reemergent, more difficult to treat, tumor. Therefore, understanding the dynamics of evolution and how to predict the future course of evolution would improve our ability to design effective treatments for cancer.

Even non-biological processes are driven by evolution and obey many of the

same laws that Darwin first laid out 150 years ago. Both languages and economies undergo evolution, driven by the same math as cancer or the origin of species. At this point it is useful to make a distinction between biological and non-biological evolution. The reason for doing so lies in the approaches that can be taken to investigate each type of evolutionary system. We know a great deal more about biological organisms than that they merely evolve. The past 50 years has resulted in an explosion of understanding of the chemistry of life and the operation of the molecular systems which compose cells. On the other hand, systems such as language and economy have, as their atomic components, humans and the human mind. While we understand much about the human mind, our grasp of its elementary functioning still pales in comparison to recent advances in biochemistry and molecular biology.

For this reason, while non-biological systems can be studied using an outside-in approach just as easily as biological systems, biological evolving systems present to us a unique opportunity to attempt to understand the mechanisms of evolution from the inside. Specifically, with biology we can explore the internal feedback mechanism which drives evolution: the "Central Dogma" of biology. This is the pathway by which information flows from an organisms nucleic acids, where it is stored, to the organisms proteins, where the information drives fundamental biochemical processes. These biochemical processes are what is eventually selected for in the process of natural selection, determining what information gets propagated, but these biochemical processes are also what does the propagation of that information.

To understand why it is so difficult to make predictions about the future dynamics of evolutionary systems, it helps to the details of how such systems works. Darwin's essential observations can be summed up in two important concepts: reproductive success and descent with modification.

Chapter 2

Design of the Polymerase Evolution Model

This chapter provides a detailed description of the design of the model system used to investigate the question of the evolution of polymerase polarity. The goal of this model is to have organisms containing either a $5' \rightarrow 3'$ or a $3' \rightarrow 5'$ nucleic acid polymerase compete with each other in order to see which strategies are either evolutionarily dominant or evolutionarily stable. The model is also designed such that the influence of temperature on the outcome of this competition can be investigated.

Overview

In simplest terms, this toy model consists of a number of individual organisms competing with each other in an environment. A number of generalizations and assumptions have been made in order to render the problem being investigated tractible, but a consistent effort has been made to remain as true to life as possible. The motivation behind this is that, while the exact values generated by this model may not be precisly those that may be observed in the laboratory, the trends observed should hold true when transfered to the bench.

The model can be largely divided up into, and thought about most clearly as, four interacting pieces. These are the environment, the organisms, the genomes, and the polymerases. For each piece decisions have been made regarding which aspects of the component are explored in depth, and which aspects are neglected, with an eye to the larger goal of investigating the role of polymerase polarity. Following is a detailed look at each component, including justifications for each decision made in the design.

Environment

A key defining feature of the model is that the environment is constrained in some ways, but not in others. Specifically, the number of organisms that can simultaneously co-exist has a hard numerical limit; a carrying capacity. On the other hand, the amount of available energy, the quantity of activated nucleotide triphosphates, and the other raw materials required for forming a cell are all considered unlimited.

In order to mimic, at least empirically, the sort of density dependant inhibition to growth that is observed in all cells as the approach the carrying capacity of their environment, a random death probability is introduced to the environment. In keeping with observations that the pressure of density dependant inhibition is greater as the population of an environment approaches the carrying capacity, the death probability is calculated with an inverse function of the remaining capacity:

Chapter 3

Experimental Results of Polymerase Modeling

Appendix A

Source Code

```
1   # Copyright (c) 2009 Joshua Ballanco
2   #
3   # class Environment
4   #
5   # Abstract: The Environment class "contains" the entire simulation. It is
6   # primarily responsible for tracking the component organisms, removing dead
7   # organisms, and stepping each organism at each time step of the simulation.
8
9   # The GenomeForEnvironment class is used to seed the starting population of
10  # the environment. It contains a +Genome+ object and the frequency at which
11  # that genome should exist in the starting population.
12  GenomeForSpecies = Struct.new(:genome, :population_frequency)
13
14  class Environment
15    attr_reader :temperature
16
17    # The +Environment+ must be initialized with the size of the initial
18         # population, the temperature of the simulation (in units of h-bond energy),
19         # and an array of structs describing the genomes to be use in creating the
20         # initial organisms.
21    def initialize(temperature, max_population, starting_population, *genomes_for_environment)
22
23               # The population frequency of all the genomes must add to 1
24      unless genomes_for_environment.inject(0.0) {|total, gfe|
25                     BigDecimal(String(total + gfe.population_frequency))
26             } == 1
27        raise ArgumentError, "population␣frequencies␣of␣genomes␣must␣total␣1"
28      end
29
30      if starting_population > max_population
31        raise ArgumentError, "the␣starting␣population␣must␣be␣less␣than␣the
32  ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
33      end
```

```
34
35       @temperature = temperature
36       @max_population = max_population
37       @organisms = []
38       genomes_for_environment.each do |genome_for_species|
39         (starting_population * genome_for_species.population_frequency).round.times do
40           @organisms << Organism.new(genome_for_species.genome.dup, self)
41         end
42       end
43     end
44
45     # Runs the environment for _max_iterations_ rounds (default is 1000).
46     def run(iterations=1000)
47       iterations.times { step }
48     end
49
50     # Set the number of threads to use, if we want to run the simulation
51         # threaded.
52     def use_threads(num_threads)
53       @num_threads = num_threads
54     end
55
56     # Before stepping the environment, calculate the probability that any
57         # individual organism will die due to resource constraints. This is modeled
58         # as a agregate probability of $\frac{1}{(N-n)+1}$, evenly distributed over
59         # the organisms in the environment, where $N$ is the carrying capacity of
60         # the environment (_max_population_) and $n$ is the number of organisms
61         # currently in the environment. At each step, the organism will either die
62         # and return nil or step and return self. At the end of stepping each
63         # organism, we compact the array to remove dead organisms.
64     def step
65       if (@num_threads && @num_threads > 1)
66         @organisms.threadify(@num_threads) do |organism|
67           organism.step
68         end
69       else
70         @organisms.each do |organism|
71           organism.step
72         end
```

```
73        end
74
75        # This is the probability that 1 organism will die:
76        death_expect = 1.0 / ((@max_population - @organisms.length) + 1)
77        if rand < death_expect
78          @organisms.delete_at(rand(@organisms.length))
79        end
80      end
81
82      # The _add_organism_ method attempts to add an organism to the environment. If there adequate capa
83      # organism is added and the method returns +true+. If the environment is currently full, then not
84      # and the method returns +false+.
85      def add_organism(organism)
86        if @organisms.length < @max_population
87          @organisms << organism
88          return true
89        else
90          return false
91        end
92      end
93
94      # The _report_ method returns a hash containing the values for this environment as well as the re
95      # iterating over the @organisms array and calling each organism's _report_ method in turn.
96      def report
97        { :temperature => @temperature,
98          :max_population => @max_population,
99          :current_population => @organisms.length,
100         :organisms => @organisms.collect{|organism| organism.report} }
101     end
102   end
```

Vita

Joshua Ballanco

Place of birth	Hazel Crest, IL
Date of birth	October 5, 1980
Education	Stevens Institute of Technology, Hoboken, NJ
	Doctoral Candidate in Chemical Biology
	expected date of graduation, September 2009

Stevens Institute of Technology, Hoboken, NJ
Masters of Science in Chemistry
May 2008

Stevens Institute of Technology, Hoboken, NJ
Bachelors of Science in Chemistry with Honors
May 2002