Modeling a *Hox* Gene Network

Stochastic Simulation with Experimental Perturbation

Thesis by
Jason Kastner

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

California Institute of Technology
Pasadena, California

2003
(Defended September 25, 2002)

Acknowledgments

Even though my name is on the title page, I am deeply indebted to a number of people, and without their help this work would not have been possible. My advisors Jerry Solomon and Scott Fraser were both instrumental in every aspect of my research, and this thesis would not have been nearly as interesting or complete without their continual guidance and help. Thanks as well to the rest of my committee, Joel Franklin, Niles Pierce, and Dan Meiron, and the funding from the Computation Molecular Biology program at Caltech, made possible by the Burroughs Wellcome fund.

All of the members of the Fraser lab helped my research, but Rusty Lansford, Paul Kulesa, Helen McBride and Reinhard Koester deserve special thanks for their advice and support. At the Stowers Institute for Medical Research, thanks to Heather Marshall, Kristen Correia, and especially Robb Krumlauf, who was incredibly generous with his time and resources.

My parents Victoria and George Kastner never failed to profess their belief in both my abilities and me, and for that I am forever indebted. But my deepest gratitude goes to my two closest friends, Jennifer Dooley and Tri Lindhom. They both finished their dissertations several years ago but were forced to relive it all again through me. Their constant support and encouragement through it all was invaluable. Thank you both.

Abstract

The *Hox* genes show a striking segment specific pattern of expression in a variety of vertebrate embryos, and have been the topic of many experimental analyses. There are now sufficient data to construct a higher-level model for the interaction and regulation of the *Hox* genes. This thesis presents the results of an investigation into a regulatory network for the early *Hox* genes. Instead of using conventional differential equation approaches for analyzing the system, a stochastic simulation algorithm has been employed to model the network. The model can track the behavior of each component of a biochemical pathway and produce computerized movies of the time evolution of the system that is a result of the dynamic interplay of these various components. The simulation is able to reproduce key features of the wild-type pattern of gene expression, and *in silico* experiments yield results similar to their corresponding *in vivo* experiments. This work shows the utility of using stochastic methods to model biochemical networks and expands the stochastic simulation algorithm methodology to work in multi-cellular systems. In addition, the model has suggested several predictions that can be tested i*n vivo*.

A tight connection was also created between the modeling and laboratory experiments. To investigate a connection between two components of the network, retinoic acid (RA) and *Hoxa1*, a novel laboratory experiment was performed to perturb the system. An RA soaked bead was implanted into the neural tube of a developing chick embryo and the effect of the exogenous RA was assayed with an *in situ* hybridization for the gene *Hoxa1*. The resulting expression patterns suggested that one aspect of the model

design was not accurate, and based on these results the model was modified to encompass the new data, without losing the fit to the original data sets. The thesis work was therefore brought full circle, thus showing the utility of an interconnected effort: the act of constructing and using the model identified interesting biology questions, and the answer to one of those questions was used to enhance the model.

Table of Contents

List of Figures and Tables

Chapter 1: Overview

> Every attempt to employ mathematical methods in the study of biological questions must be considered profoundly irrational and contrary to the spirit of biology.
>
> If mathematical analysis should ever hold a prominent place in biology—an aberration which is happily almost impossible—it would occasion a rapid and widespread degeneration of that science.
>
> - Auguste Comte, 1871

Introduction

Every applied and computational mathematics thesis should start with a physical problem, and in that respect this thesis is true to form. Instead of culling a problem from physics however–the traditional inspiration for much of applied mathematics–the problem under investigation in this work was drawn from developmental biology. The goal of this thesis was to investigate a relevant and interesting biological problem from both the modeling and experimental arenas, and show the efficacy of an interconnected effort. This thesis presents the results of an investigation into a regulatory network for a set of genes expressed in the developing brain, the *Hox* genes. The network was created through integrating the results of numerous biology papers and constructing a higher-level model for the interaction and regulation of the *Hox* genes in a multicellular context.

Instead of using conventional differential equation approaches for modeling the resulting system, a stochastic simulation algorithm (SSA) has been employed to model

the network. This work improves on previous SSA investigations that had been limited to intracellular systems by expanding the SSA to work in an intercellular arena. One of the troublesome problems with modeling a multi-cellular system involved cell synchronization, and this was solved with the use of a priority queue to time-order the cells. The model tracks the behavior of each component of a biochemical pathway and captures the dynamic interplay of the various components in the multi-cellular system. The data can be rendered as computerized movies of the time evolution of the system. The simulation is able to reproduce key features of the wild-type pattern of gene expression, and *in silico* experiments yield results similar to their corresponding *in vivo* experiments. In addition, the model has suggested several predictions that can be tested *in vivo*.

An important goal of this thesis was a tight connection between the modeling and experimental work, and two novel perturbation experiments aimed at testing components of the model network were designed. The first investigation addressed the connection between two genes in the network, *Hoxb1* and *Krox20,* and the published hypothesis that *Krox20* is repressed by *Hoxb1* expression (Barrow et al., 2000). A specially constructed piece of DNA designed to repress *Hoxb1* was introduced into young chick embryos, and the effect on *Krox20* expression was assayed. The DNA did not, however, appear to work as intended. The second experiment explored the connection between retinoic acid and *Hoxa1* by altering the normal retinoic acid distribution in the embryo. This was accomplished by implanting a retinoic acid soaked bead into the midbrain of a developing chick and assaying the expression of *Hoxa1*. This experiment yielded intriguing results, and the resulting data suggested that one aspect of the model design

was not accurate. Based on these results the model was modified to encompass the new data, without losing the fit to the original data set. The thesis work was therefore brought full circle, thus showing the utility of an interconnected effort: the act of constructing the model identified interesting biology questions, and the answer to one of those questions was used to enhance the model.

Interdisciplinary Work

With such a strong focus on interdisciplinary research, this work presented a number of challenges that are not typically found in a conventional thesis. They started with the need to learn the vocabulary of a new field. This was accomplished by sitting in on biology courses, reading the biology literature, and interacting with people working in a biology laboratory. At the same time, a search to identify a tractable yet interesting problem was undertaken. The prospect of modeling a gene network appeared fairly early in the research process, yet it took a great deal of time to identify a particular network.

The molecular studies of the hindbrain have offered sufficient details to assemble a model for the interactions important in regional control of gene expression. These factors helped identify a system in which to work; the interconnection of the early *Hox* genes and their connection to retinoic acid. The direct coupling of the stochastic simulation algorithm implementation of a network and individual molecular events would seem to lend itself to both the analysis and logical organization of the ever growing data on the control of *Hox* genes in the developing hindbrain.

One of the important features of the *Hox* system is that the amount of molecular information that has been gathered about the regulatory mechanisms allows for a

synthesis and construction of a higher-level system of interaction. At the same time, the data is far from complete, thus leading to questions that can be investigated through simulation. These include investigations of hypothesized interactions, mechanisms of interaction, and perturbations of the system.

Another key feature of the *Hox* network was an animal model, the chick hindbrain, which allowed for experimental perturbation of the system *in vivo*. A carefully designed experiment could be connected back to the model, and the data gathered from the experiments would offer support for, or evidence against, model hypotheses.

Finally, research into the *Hox* genes is relevant because of their strong connection to diseases. There is evidence linking *Hox* family members to leukemia (Thorsteinsdottir et al., 2001) and breast cancer (Lewis, 2000), and connections to genetic diseases include obsessive-compulsive disorder (Greer, 2002) and autism (Ingram et al., 2000; Rodier, 2000).

The laboratory work was designed from the outset to be a crucial part of this research. The experiments are intimately related to the *Hox* network, and early on in the work it was necessary to move beyond the literature and start work in a laboratory. The literature and consultations with experimentalists provided the initial guidance in perturbation techniques—the bead implantation (Chapter 4) and electroporation (Appendix A)—but the refinement of the methods came through trial and error. To do these experiments, it was necessary to learn an array of supporting techniques. These included early chick embryology and development, tissue culture, microscopy, and a

number of molecular biology techniques including antibody staining, cloning, and *in situ* hybridization. Many of these techniques are described in the thesis. During the course of working in the laboratory, numerous problems that are never mentioned in the literature or classes appeared on an almost daily basis. The Vital Stain experiment in Chapter 4 is an illustrative example.

To present this interdisciplinary work in the proper context, the thesis is broken into the following 5 chapters: Chapter 1 provides an overview of modeling biological problems, an introduction to modeling gene networks, as well as some comments about the goals of modeling in general. Chapter 2 focuses on the modeling of enzyme kinetics by presenting stochastic and deterministic implementations of the basic enzyme reaction and a comparison of the two. Chapter 3 includes an introduction to both developmental biology and the specific biology of the system under investigation. It goes on to present the model itself, and a sensitivity analysis of the model. Chapter 4 is devoted to experimental results, and how the experiments described tie back into the model. Chapter 5 contains the summary and a discussion of the work. The Appendices contain more experimental results, the source code for the simulations, and the laboratory protocols used to perform the experiments.

Biological Modeling

Over 170 years after Comte made his thoughts concerning the role of mathematics in biology known, his sentiments are perhaps too widely shared in the biology community. D'arcy Wentworth Thompson echoed Comte's sentiment when he remarked

that “The introduction of mathematical concepts into natural science has seemed to many men no mere stumbling-block, but a very parting of ways” (Thompson, 1942).

Practically speaking, the reasons for the schism between math and biology are many. They start with the language barrier, a common obstacle between many fields. Unlike math and physics, which are inextricably linked by their vocabulary, math and biology each have a vocabulary that is very difficult for the outsider to understand. This has created a climate that does not encourage true interdisciplinary work and there are numerous instances of mathematics used to solve problems that are supposedly biological in nature, but in truth have little connection. The language barrier also presents problems when communicating the results of the work, but it has been shown that publishing the research in a journal relevant to the new field is an effective form of interdisciplinary information transfer (Pierce, 1999). Therefore, the fact that a portion of this work has been published in the journal *Developmental Biology* (Kastner et al., 2002) is a notable achievement.

Another problem is that modeling biological processes is inherently difficult; there are relatively few “toy problems” that can be easily identified, extracted, and solved. This often leaves an investigator in the difficult position of trying to model a system before it is well characterized. It is sometimes suggested that all the parts of the system must be known before a model can be created, or that any potential modeling approach must be proved on the simplest system before trying to apply it to something more complex. These objections are sometimes put forth as reasons not to start work on a problem, but they are shortsighted and in truth much can be accomplished by trying to model even poorly characterized biological problems. Indeed, a central reason for

modeling biology using mathematics and computers is precisely because the biological systems are so incredibly complex. The facts of the matter are simply these: all the parts of any real biological system are likely to never be known, and even the simplest biological systems are more complex than can be handled by any supercomputer. To quote an oft-repeated sentiment during many biology lectures: "but it's more complicated than that." Not only is it more complicated than that, it is more complicated than we can begin to imagine. Therefore, a major part of the problem with biological modeling is finding tractable yet interesting problems.

Finally, the scientific community is still trying to develop a mathematical framework for biological problems. There is no $F = ma$ for biology, and a variety of techniques can often be employed for each problem that appears. The closest biology has come to a universal law is the Central Dogma which states that genetic information is carried on DNA, then transcribed to RNA and subsequently translated to proteins. Adding to this problem is that data arising from biology experiments, especially in developmental biology, are often qualitative and don't always lend themselves to a rigorous mathematical analysis.

Despite these objections, it is important to try to bring communities together as there is much they can offer each other. For the mathematicians, biology affords a relatively untapped spring of interesting problems, and the opportunity to shape the future direction of investigations. For the biologists, mathematics can provide a framework for the biology problems, especially considering the sheer amount of biology data being generated. It can also be used to quantify results and suggest experiments to test hypotheses, ultimately adding to the understanding of how the biology may work.

Gene Networks

One focus of traditional biology examines single genes or proteins in isolation. While this provides vital information, it is the interaction of these pieces that provides biological results. The logical next step is therefore combining the data from various sources to build a hierarchal picture of the true interactions of the pieces of the pathways. Because of the deluge of information, computer models are the key to the future of the information integration and to the understanding of how the systems work. Not only that, but by a thoughtful investigation into a system, it is even possible to determine the part of the model which may be missing or is not well understood. An excellent example of this has recently appeared with the use of a model to discover a missing control module for a sea urchin gene (Yuh et al., 2001).

Biological networks are the collection of biochemical entities (including messenger RNA, proteins, DNA, ions, or other molecules, like hormones), which interact to produce biological results. An analysis of these systems seeks to elucidate information about the interactions between the genes and their derivatives, and also hopes to provide predictive results about the overall behavior of the system. This type of work is commonly called systems biology because it seeks to simultaneously study the complex interaction of many levels of biological information.

Genetic networks currently lie in the forefront of biological research, and are in the border area where computer simulations and molecular biology meet. The most successful efforts have tightly coupled the modeling and experimental efforts (*cf.* Yuh et al., 1998; Yuh et al., 2001). They are also an area of increasing interest, evidenced by the

growth in the literature. Five years ago a literature search on the term "gene network" returned only 3 references, and none of the works involved modeling. In the first nine months of 2002 however, the same search produced nine times as many results, and a dozen of them clearly involve modeling of some sort.

Various methods have been employed to model biological networks including Bayesian networks (Friedman et al., 2000), rule based formalisms (Meyers and Friedland, 1984), true Boolean systems (Kauffman, 1993) and Boolean/continuous hybrids (Yuh et al., 1998; Yuh et al., 2001) but ordinary differential equations have been the preferred method to construct and analyze biochemical network models. Using the Law of Mass Action, which states that the rate of the reaction is proportional to the concentration of the reactants, it is possible to write down a set of coupled differential equations that hope to describe the time evolution of the system. The reasons for the prevalence of mass action based kinetic analysis are many, but by far the most important one is that the approaches based on differential equations produce results that are in general in good agreement with the data (*cf.* Hynne et al., 2001; Poolman et al., 2001). In addition, differential equations come with a wide range of analysis tools that allow for a detailed investigation of the model properties. But as will be addressed in Chapter 2, differential equations may not be appropriate for modeling biological processes in the small volumes inherent in single living cells.

Compared to differential equations, and despite their prevalence in modeling pure chemical processes, stochastic approaches in biology are still in a relative infancy. This is currently changing, and generalized tools for constructing and analyzing stochastic simulations are now starting to appear (Bray et al., 2001; Kierzek, 2002). A stochastic

process is one governed by a random process, and in a biological context this means that the system is subject to fluctuations. These fluctuations could be in the number of molecules present, the time it takes for a molecular creation or decay process, or the length of time molecules are bound together. More attention has been focused lately on stochastic effects in biology, especially as evidence shows that stochastic effects play major roles in gene expression (Greenwald, 1998; Ko, 1992; Zlokarnik et al., 1998). Instead of treating these factors explicitly, some differential equation approaches attempt to capture stochastic effects by adding a "noise" term to their otherwise deterministic treatment (*cf.* Meinhardt and de Boer, 2001). The resulting "ordinary" differential equation is called the Langevin equation and is of the form

$$\frac{dX(t)}{dt} = -aX(t) + f(t) \tag{1.1}$$

where the noise function $f(t)$ is assumed to be Gaussian and delta-correlated. But in effect this makes the noise term just another parameter instead of capturing it in a physical meaningful way. This may be a somewhat misguided approach: if there are fluctuations in the system that need to be accounted for, it might be preferable to incorporate those effects at the beginning in a way that is physically intuitive and physically based.

Stochastic Simulation

As opposed to the deterministic view in which the reaction constants are the rates, reaction constants in the stochastic approach are considered to describe the probability (per unit time) that a reaction occurs. With this formulation, the chemical system can be

thought of a Markovian random walk in the space of the reacting molecular species. The time evolution of the system is described by the solution of a single differential difference equation, often called the master equation. The independent variables of the master equation are time and the populations of the reacting species. The master equation can be transformed into a partial differential equation by the use of a generating function.

From a mathematical point of view, the set of equations resulting from the Law of Mass Action is usually easier to solve than the corresponding master equation or the associated partial differential equation. In reality, it turns out that if the system involves more than a few reactants and chemical reactions, an analytic solution is out of reach for either method, and it is necessary to use a numerical scheme (McQuarrie, 1967). Of course numerical methods for solving even a single partial differential equation can be a research topic in and of itself; instead what was really needed was a general method for attacking the master equation. This came in 1976 when Dan Gillespie introduced the stochastic simulation algorithm, described in the next chapter (Gillespie, 1976).

Adam Arkin appears to be the first to use Gillespie's method in a biological context with a study of the growth of phage λ, a virus that infects the bacteria *E. coli* (Arkin et al., 1998; McAdams and Arkin, 1998). This thesis shows that stochastic simulation has a much wider range of applications by applying the methodology to a larger system, namely a collection of cells, each with a much more complicated network containing more molecular species than phage λ.

A Caveat Concerning Modeling

With all these attempts to model a biological system, it is important to keep track of the goals and the pitfalls of modeling in general. This is most succinctly put in an article concerning the nature of numerical modeling in the earth sciences, but the nature of the arguments apply to any field in which models are created.

> Verification and validation of numerical models of natural systems is impossible. This is because natural systems are never closed and because model results are always nonunique. Models can be confirmed by the demonstration of agreement between observation and prediction, but confirmation is inherently partial. Complete confirmation is logically precluded by the fallacy of affirming the consequent and by incomplete access to natural phenomena. Models can only be evaluated in relative terms, and their predictive value is always open to question. The primary value of models is heuristic. (Oreskes et al., 1994)

This situation is clearly illustrated in this thesis. The *Hox* network model was constructed using the relevant biochemistry and biology, and the model results were in good agreement with the published laboratory experiments. When a new experiment was performed to test an implementation decision of the model, it turned out that the model was not in agreement with the new experimental results. This resulted in a change to the model to fit the new experimental data, but the new simulation results were essentially indistinguishable from the original results. So while the new model must now be seen as better, in so far as it is consistent with more of the real data, there is unfortunately no guarantee that future predictions will match laboratory observations more closely. This is especially true given the incredibly dynamic nature of the system and the model.

Of course, these criticisms are valid for any model that seeks to describe a natural system, and so it is important to remember what models actually can do: they are useful in identifying parts of a problem that are in need of further study, and in identifying the

data that is relevant to the problem at hand. Furthermore, the very act of constructing a model can stimulate questions about how the natural system behaves. In this instance, the questions lead to the retinoic acid soaked bead experiment described in Chapter 4. The resulting data adds to the understanding of the connection between retinoic acid and the gene *Hoxa1*, in particular, and the network of genes patterning the brain in general.

References for Chapter 1

Arkin, A., Ross, J., and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* **149,** 1633-48.

Barrow, J. R., Stadler, H. S., and Capecchi, M. R. (2000). Roles of Hoxa1 and Hoxa2 in patterning the early hindbrain of the mouse. *Development* **127,** 933-44.

Bray, D., Firth, C., Le Novere, N., and Shimizu, T. (2001). StochSim.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayseian networks to analyze expression data. *J. Comput. Biol.* **7,** 601-620.

Gillespie, D. T. (1976). A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *Journal of Computational Physics* **22,** 403.

Greenwald, I. (1998). LIN-12/Notch signaling: lessons from worms and flies. *Genes Dev* **12,** 1751-62.

Greer, J. M. a. C., M.R. (2002). *Hoxb8* Is Required for Normal Grooming Behavior in Mice. *Neoron* **33,** 23-34.

Hynne, F., Danø, S., and Sørensen, P. G. (2001). Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophysical Chemistry* **94,** 121-163.

Ingram, J. L., Stodgell, C. J., Hyman, S. L., Figlewicz, D. A., Weitkamp, L. R., and Rodier, P. M. (2000). Discovery of allelic variants of HOXA1 and HOXB1: genetic susceptibility to autism spectrum disorders. *Teratology* **62,** 393-405.

Kastner, J. C., Solomon, J. E., and Fraser, S. E. (2002). Modeling a Hox Gene network *in silico* using a Stochastic Simlulation Algorithm. *Developmental Biology* **246,** 122-131.

Kauffman, S. A. (1993). "The Origins of Order." Oxford University Press, Oxford.

Kierzek, A. M. (2002). STOCKS: STOChastic Kinetic Simulations of biochemical systems with gillespie algorithm. *Bioinformatics* **18,** 470-481.

Ko, M. S. (1992). Induction mechanism of a single gene molecule: stochastic or deterministic? *Bioessays* **14,** 341-6.

Lewis, M. T. (2000). Homeobox genes in mammary gland development and neoplasia. *Breast Cancer Res*. **2,** 158-169.

McAdams, H. H., and Arkin, A. (1998). Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct*. **27,** 199-224.

McQuarrie, D. A. (1967). Stochastic Approach to Chemical Kinetics. *Journal of Applied Probability* **4,** 413-478.

Meinhardt, H., and de Boer, P. A. J. (2001). Pattern formation in *Escherichia coli*: A model for the pole-to-pole oscillations of Min proteins and the localization of the division site. *PNAS* **98,** 14202-14207.

Meyers, S., and Friedland, P. (1984). Knowledge-based simulation of genetic regulation in bacteriophage lambda. *Nucleic Acids Res*. **12,** 1-9.

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* **263,** 641-646.

Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. *Journal of the American Society for Information Science* **50,** 271-279.

Poolman, M. G., Ölçer, H., Lloyd, J. C., Raines, C. A., and Fell, D. A. (2001). Computer modelling and experimental evidence for two steady states in the photosynthetic Calvin cycle. *Eur. J. Biochem*. **268,** 2810-2816.

Rodier, P. M. (2000). The early origins of autism. *Sci. Am*. **282,** 56-63.

Thompson, D. A. W. (1942). "On Growth and Form." Dover, Mineola, New York.

Thorsteinsdottir, U., Kroon, E., Jerome, L., Blasi, F., and Sauvageau, G. (2001). Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia. *Mol. Cell. Biol*. **21,** 224-234.

Yuh, C. H., Bolouri, H., and Davidson, E. H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279,** 1896-902.

Yuh, C. H., Bolouri, H., and Davidson, E. H. (2001). Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128,** 617-29.

Zlokarnik, G., Negulescu, P. A., Knapp, T. E., Mere, L., Burres, N., Feng, L., Whitney, M., Roemer, K., and Tsien, R. Y. (1998). Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science* **279,** 84-8.