

# The Perfect *C. elegans* Project: An Initial Report

**Hiroaki Kitano**

Sony Computer Science Laboratory  
3-14-13 Higashi-Gotanda, Shinagawa-ku  
Tokyo 141, Japan  
kitano@csl.sony.co.jp

**Jun Kitazawa**

Dept. of Media and Governance  
Keio University  
Fujisawa, Japan

**Shugo Hamahashi**

Dept. of Electrical Engineering  
Keio University  
Yokohama, Japan  
shugo@mt.cs.keio.ac

**Sean Luke**

Department of Computer Science  
University of Maryland  
College Park, MD 20742 USA  
seanl@cs.umd.edu

## Abstract

The soil nematode *Caenorhabditis Elegans* (*C. elegans*) is the most investigated of all multi-cellular organisms. Since the proposal to use it as a model organism, a series of research projects have been undertaken, investigating a wide range of on various aspects of this organism. As a result, the complete cell lineage, neural circuitry, and various genes and their functions have been identified. The complete *C. elegans* DNA sequencing and gene expression mapping for each cell at different times during embryogenesis will be identified in a few years. Given the abundance of collected data, we believe that the time is ripe to introduce synthetic models of *C. elegans* to further enhance our understanding of the underlying principles of its development and behavior. For this reason, we have started the *Perfect C. elegans Project*, which aims to ultimately produce a complete synthetic model of *C. elegans* cellular structure and function. This paper describes the goal, the approach, and the initial results of the project.

## 1 Introduction

When Sydney Brenner proposed the investigating *C. elegans* to the Medical Research Council, he chose it because it was the simplest of all differentiated organisms [Brenner, 63]. The decision proved wise, generating a number of fruitful results, including the complete identification of cell lineage of the worm [Sulston and Horvitz, 77, Kimble and Hirsh, 79, Sulston et al., 83] and its full neural circuit topology including all synapses and gap junctions [White et al., 86]. Using a whole mount *in situ* hybridization, the complete *C. elegans* DNA sequencing and expression pattern mapping are now within our reach [Sulston et al., 92, Tabara et al., 96]. These investigations clarify what *C. elegans* is composed of and how each of its individual components work. However, an under-

standing of the components and their isolated functions do not lead to an understanding of the dynamics behind the global development and behavior of this organism. Despite the fact it is the simplest known differentiated organism, *C. elegans* dynamics are still too complex for us to fully grasp at this time.

This situation typifies the problems that lie ahead in molecular biology. The human genome projects and other genome sequence projects will soon identify complete DNA sequences and clone all genes, giving the ultimate reductionist view of ourselves and other organisms; however, this does not directly lead to an understanding of the full interactions that make up living organisms. Since many phenomena occurring in living organisms are complex and non-linear, it is almost impossible to fully understand their characteristics without the aid of simulation and modelling. Thus, we claim that an in-depth understanding may be only be achievable through reconstructing the phenomena with computer simulation.

In a first attempt at this approach, we have developed detailed simulation models of *C. elegans* and *Drosophila melanogaster* (another widely studied organism). Given the abundance of data and the expected progress of analytical methods for these organisms, they're the natural target for this approach.

## 2 *Caenorhabditis elegans*

*C. elegans* is a small worm found in soil and ubiquitously observed throughout the world. It accounts for the largest biomass on earth. It has a life span of about 3 days and feeds on bacteria. *C. elegans* has no female sex (only hermaphrodites and males), so most every worm in the population is genetic clone. This, plus the simple nature of the organism, means that lineages, positions, and interactions of *every single cell* can be mapped out. The adult male *C. elegans* has exactly 1031 somatic nuclei, and the adult hermaphrodite has 959

somatic nuclei. The lineage of these cells has been fully identified by the extensive work of various research groups (one in particular: [Sulston et al., 83]).

The nervous system of *C. elegans* is relatively simple. Its hermaphrodites have only 302 neurons and 56 glial and associated support cells. This accounts for 37% of all somatic cells. In the adult male, the number of neurons is 381, and there are 92 glial and support cells, which is 46% of all somatic cells. White differentiated these neurons into 118 classes and reported that there are about 5,000 chemical synapses, 2,000 neuromuscular junctions, and 600 gap junctions [White et al., 86].

A haploid *C. elegans* genome is composed of  $8 \times 10^7$  nucleotide pairs. The *C. elegans* genome project is being carried out by the Sanger Center and Washington University, which has already sequenced more than 25% of its genomes. Current progress suggests that all genes, estimated to number about 13,000, will be identified within a few years.

The mechanism of fate determination involving maternal genes has been intensively investigated. However, fate determination in later cells is largely unknown because downstream genes have not been identified. In order to investigate the genetic interactions for fate determination, a project to identify the genes that are expressed in a specific cell lineage has been initiated at the National Institute for Genetics [Tabara et al., 96]. This project uses *in situ* hybridization on whole mount embryos to identify the expression of genes at specific cells at specific times during embryogenesis.

Many mutants have been isolated which can be used for genetic analysis. For example, the *ced* family affects programmed cell death, and mutations in the *lin* family of genes cause cell lineage abnormality. There is a large list of genes and their phenotypical disorders, allowing for a wide range of manipulations to help in the investigation of cellular development. Other manipulations are possible during embryogenesis through laser ablation or direct micro-manipulations.

### 3 Project Goals

Given these biological accomplishments and on-going efforts, the Perfect *C. elegans* Project aims to create a detailed simulation model of *C. elegans* to promote our understanding of the organism and of life in general.

By implementing a detailed model, we can verify how much we really do understand about *C. elegans*. Due to complexities of genetic and cell to cell interactions, and the large number of participating genes and their products, it is extremely difficult for us to fully understand and verify whether a certain hypothesis can be supported by biological findings. A possible use of the simulation model is to bring the genetic interactions of genes into focus to see if hypothetical genetic interactions can consistently recreate phenotypical changes that match known biological observations. If the results of simulation differ from actual observations, there are two possibilities: either

the simulation or the hypothesis is incorrect. The accuracy of simulations needs to be carefully examined and verified before the model can serve as a testbed for hypotheses. It should be noted, however, that even with a limited accuracy model, hypotheses can be tested if they focus on specific aspects of genetic interactions and involve logical interrelationships of interactions rather than details concerning the subtle balance of participating components.

By the same token, a well designed model can be used for predicting the workings of possible molecular mechanisms which are involved in specific phenomena. The ideal approach is for the model to provide a set of possible molecular mechanisms for specific phenomena whose true mechanisms are unknown. Experimental biologists can then design experiments to identify which of these mechanisms actually exists.

Implementing the model can be extremely useful for biologists studying *C. elegans*. It can be a comprehensive visual database of the organism, and assist in cell identification. Currently, the most widely used computer assisted system for *C. elegans* is the Angler system, also called the 4D system, developed by the Sanger Center. Angler is essentially a collection of tagged images taken by Nomarski confocal microscopy. However, the Angler system does not provide the capability to rotate images or to animate the embryogenesis process. A good computer model can compliment the Angler system by providing computer graphic images and simulations linked with a set of optical images.

While modeling *C. elegans* on a computer is useful in its own right for augmenting biological knowledge, the visualization of embryogenesis is one of the most persuasive tasks of the project. Visualization helps researchers to conceptualize both global and local issues in embryogenesis, and to identify cells during experiments. This requires tools to assist biological observations using three-dimensional computer graphics of the complete *C. elegans* development process (possibly coupled with image understanding techniques for automatic data acquisition and semi-automatic cell identification), a simulator for a complete neural circuit, and an integrated database on *C. elegans*.

To begin meeting these goals, we have so far developed three modelling and visualization systems for *C. elegans* cellular data. The first, a embryogenesis visualization system, uses a sophisticated dynamics model to display the embryogenesis process cell-by-cell. The second, a neural simulation system, models the known neural circuitry involved in *C. elegans* thermotaxis. The third, a portable Java-based 3D visualizer, displays simple cell position, neural connections, and other information using new positional data recently made available, without attempting the sophisticated cellular modelling and 3D rendering of the first system.

## 4 Visualization and Simulation of Embryogenesis

As the first step toward these goals, we have developed a computer graphics system which helps visualize the development process of *C. elegans*. This visualization system is an appropriate starting point because it provides a 3D model of *C. elegans*, so that the cell to cell interaction dynamics, at both the physical and chemical levels, can be superimposed on the current model. This could greatly help research in developmental biology. The current system is based on the cell lineage and cell location data published in [Sulston et al., 83]. The system generates a computer graphics image from the division of the first cell to around 600 minutes after the first cell division.

It is no trivial task to create a reasonably accurate computer graphics image based on the Sulston data because the information necessary to create a full three dimensional model is missing. The following information was available:

- the complete cell lineage chart
- hand-drawn pictures in 2-1/2 dimensions
  - all 28 cells at 100 minutes
  - 55 out of 180 cells at 200 minutes
  - 137 out of more than 350 cells at 260 minutes
  - 156 out of more than 350 cells at 270 minutes
  - nearly all cells at 430 minutes
- qualitative descriptions of the embryo shape
- qualitative descriptions of the disparity in the size of divided cells
- general information on migration

In addition, the lineage chart only indicates approximate division time and a rough direction of the division, such as anterior-posterior, dorsal-ventral, or left-right. For example, the cell lineage of two ring inter-neurons (ADAL and ADAR) are given as ADAL AB.plapaaaapp and ADAR AB.prapaaaapp. This means that the ADAL cell was derived from the founder cell AB through a series of divisions: the first division was anterior-posterior, and ADAL is derived through the posterior child. The second division was left-right, and ADAL is derived through the left child. The remaining divisions are anterior (a), posterior (p), four anterior (aaaa) divisions, and two posterior divisions (pp). ADAR is the symmetrical counterpart to ADAL, and has a similar lineage but made a division to a right cell after the posterior division.

In reality, the direction of cell divisions are not necessary aligned with these axes. The simulator has to deal with this issue carefully. By the same token, the location of cells in Sulston's paper are drawn as a series of two dimensional figures. The vertical position of cells are shown only by using

circles having three levels of thickness. While the approximate two-dimensional location of cells can be estimated from these figures, it only provides a crude idea of the depth (or vertical position) of cells. Again, the simulation method needs to develop a good estimate of the true 3D location of these cells.

These efforts are necessary to make a best guess on the three dimensional position of cells. This is essential to create accurate computer graphics images. Obviously, a straightforward way would be to collect three dimensional position and shape data in a series of time steps, to create graphics based on these data. This approach is extremely labor-intensive because each cell must be manually identified at each time step. In order to collect this data, researchers need a crude 3D visualization tool to assist in the identification of cells and their position. In addition, such an approach cannot be applied to visualize the development of organisms which have only a limited amount of position information. Therefore, we need to develop a system which can generate reasonably accurate 3D computer graphics images based on limited data sets.

Our strategy to overcome this problem is to merge simulation with data. First, in order to assure the accuracy of the computer graphics image, cells must be in the position given by the observed data. Second, various simulation techniques are used to fill in missing information, such as the location of cells not provided in the data. To do this, the system computes forces between cells such as the force which pushes back colliding cells. However, if only a dynamic simulation is used to decide the position of cells, some cells will not be in the position described in the observed data because of cell movement. In order to compensate for this discrepancy, the force that a cell generates for its movement is estimated using inverse kinematic techniques and is added to the cell's force vector. Cell movements are computed as objects in a viscous fluid.

### 4.1 Data Records

A set of data obtained from actual observation must be represented in machine readable form. Table 1 shows part of data record that represents cell lineage and cell positions taken from the Sulston data. R is the record type (f and p means a full record, x means a terminal record), and  $d_1$  and  $d_2$  are the names of daughter cells. When '\*' precedes a symbol, such as '\*a', it is concatenated to the name of mother cell to create the name of the daughter cell. For example, AB's daughters are labeled '\*a' and '\*p', in the table. The full name of these daughter cells are ABa and ABp. x, y, and z indicate the direction of division. The relative volume of the cell  $d_1$  is shown under 'vol'. The time of the division is indicated under 'time'.  $x_1$ ,  $y_1$ ,  $z_1$ ,  $r_1$ , and  $t_1$  are the x, y, and z positions and cell radius at time  $t_1$  (similarly, later columns continue describing x, y and z positions and cell radii at additional time points).

This database contains most of the information provided in the Sulston paper. Even when we obtain more accurate data, we can continue to use this data record by making minor

| R | Cell name  | $d_1$ | $d_2$ | x  | y | z  | vol | time | $x_1$ | $y_1$ | $z_1$ | $r_1$ | $t_1$ | $x_2$ | $y_2$ | $z_2$ | $r_2$ | $t_2$ |
|---|------------|-------|-------|----|---|----|-----|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| f | Egg        | AB    | P1    | 0  | 1 | 0  | .6  | 0    | 0     | 0     | 0     | 5.4   | 0     |       |       |       |       |       |
| p | AB         | *a    | *p    | 0  | 1 | -1 | .5  | 18   |       |       |       |       |       |       |       |       |       |       |
| p | ABa        | *l    | *r    | -1 | 0 | 0  | .5  | 35   |       |       |       |       |       |       |       |       |       |       |
| p | ABp        | *l    | *r    | -1 | 0 | 0  | .5  | 35   |       |       |       |       |       |       |       |       |       |       |
| p | P1         | EMS   | P2    | 0  | 1 | -1 | .6  | 19   |       |       |       |       |       |       |       |       |       |       |
| p | EMS        | MS    | E     | 0  | 1 | 0  | .6  | 38   |       |       |       |       |       |       |       |       |       |       |
| p | P2         | C     | P3    | 0  | 1 | 1  | .6  | 44   |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
|   |            |       |       |    |   |    |     |      |       |       |       |       |       |       |       |       |       |       |
| p | ABalaaappr | l     | r     | -1 | 0 | 0  | .5  | 287  | 5.0   | 15.8  | -2    | 1.8   | 260   | -3.9  | 15.3  | -5    | 1.7   | 270   |

Table 1: Data record for cell lineage and cell positions.

changes. One point which needs to be changed is the way the cell division direction is specified. In the current data structure, it is represented as a three dimensional vector where each dimension has only three discrete values (1, 0, and -1). This might be changed to allow continuous values. Cell shape information is also not included in the current version, and hopefully will be added in the future.

## 4.2 Interpolation and Kinematics

Although the database contains some data on cell lineage, position, and other information, this is not sufficient for visualization. The simulation system uses interpolation and kinematics to estimate positional values when data isn't available.

When a cell divides, its two daughter cells divvy up the volume of the mother cell. The volume conservation rule imposed at division-time can help determine the approximate position of the center of each daughter cell for daughter cells. Immediately after each cell division, the new cells are moved to positions consistent with known dynamics. The position of the cell must also be consistent with the estimate provided by the observed positions of the daughter cells. Given the center-of-cell positions of the two daughter cells ( $x_1, y_1, z_1$  and  $x_2, y_2, z_2$ ), the center-of-cell position of the mother cell is assumed to be in the middle of the daughter cells ( $x = \frac{x_1+x_2}{2}, y = \frac{y_1+y_2}{2}, z = \frac{z_1+z_2}{2}$ ).

However, this assumes that there is no cell movement, a consideration which requires a more complex estimation strategy. Although the division time of each cell is known, the location of the division is not described in the Sulston paper. Thus, we had to interpolate the position where the cell division takes place. For a data, we can use the position of the mother cell (cell number = 1) at time  $t_1$  as  $\vec{x}_1(t_1)$  and the positions of the daughter cells (cell numbers 2 and 3) at time  $t_3$  as  $\vec{x}_2(t_3)$  and  $\vec{x}_3(t_3)$ . The center-of-cell position of the mother can be obtained as the middle position of the center of two daughter cells:  $\vec{x}_c = \frac{\vec{x}_2 + \vec{x}_3}{2}$ . Assuming that the division took place at time  $t_2$ , the position of cell division can be obtained by using the linear interpolation as:

$$x_1(\vec{t}_2) = A\vec{x}_1(t_1) + B\vec{x}_c(t_3) \quad (1)$$

$$A = \frac{t_3 - t_2}{t_3 - t_1} \quad (2)$$

$$B = 1 - A \quad (3)$$

This is used for simulation when less computation power is available. The spline interpolation method is used for the default simulation model:

$$\vec{x}_1(t_2) = A\vec{x}_1(t_1) + B\vec{x}_c(t_3) + C\vec{x}_1''(t_1) + D\vec{x}_c''(t_3) \quad (4)$$

$$A = \frac{t_3 - t_2}{t_3 - t_1} \quad (5)$$

$$B = 1 - A \quad (6)$$

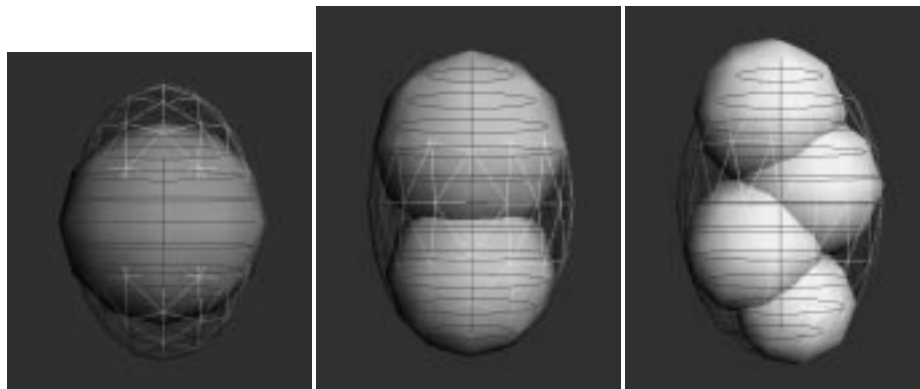
$$C = \frac{1}{6}(A^3 - A)(t_3 - t_1)^2 \quad (7)$$

$$D = \frac{1}{6}(B^3 - B)(t_3 - t_1)^2 \quad (8)$$

These are the basic techniques for interpolation based on the available data. However, because the time intervals of existing data are so large that simple interpolation does not capture the complex movement of cells, we need a simulation of dynamics as well. Such an approach is also essential to simulate phenotypical changes due to the microscopic operations of cell positioning, laser ablation, and mutant analysis where some cell movements are affected. The current version models these kinetic interactions of cells as a coupled continuous elastic system, where many masses are coupled by elastic springs. Although cells are displayed as spheres and no elasticity is visualized, it appears from the graphics image that cells are overlapping each other. In such cases, a repulsive force is imposed on overlapping cells. The simple way to approximate this force is a linear increase of force proportional to the length of the overlap. A more detailed simulation demands uses a more concrete force calculation. The details of the algorithms and equations used in these kinematics methods are beyond the focus of this paper.

### 4.2.1 Bending

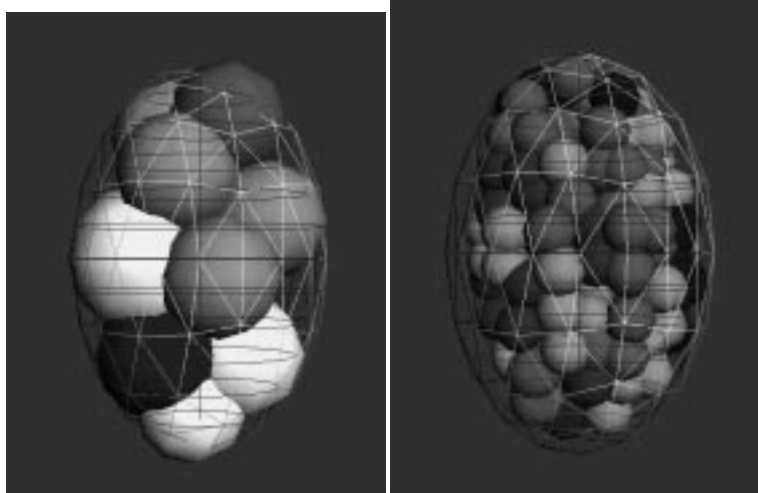
The shape of the embryo changes after 200 minutes to create the so-called "comma" or "2-fold" shape. Since the driving



(A) One cell

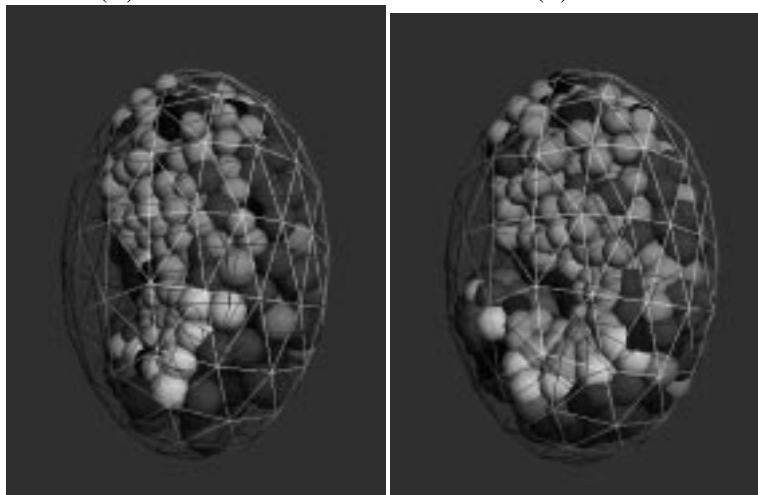
(B) Two cells

(C) Four cells



(D) 50 min.

(E) 200 min.



(F) 350 min.

(G) 400 min.

Figure 1: Snapshots of computer graphics images of *C. elegans* embryogenesis

force which bends the embryo's shape is not identified biologically, the computer simulation imposes a top-down kinematic. At 200 minutes, a cylindrical coordinate system is imposed to describe the cell locations, and the cylindrical coordinate system itself is bent slowly to agree with the observed shape so that cells change their absolute positions according to the bending of the cylindrical coordinates (Fig 1. F and G). As soon as a mechanism for bending is identified, we will incorporate it into our system, and remove this top-down part of the simulation.

### 4.3 Implementation

The current implementation of the system uses the C++ programming language, the OpenGL graphics library, and the SGI sphere library on a Silicon Graphics workstation. The simulation gives a real-time animation of the embryogenesis process. Some of the images are shown in Figure 1. Colors assigned to cells are based on the cell fate color map. A, B, and C are snapshots of the CG image at the initial stage (egg), two cell stage, and four cell stage, respectively. D, E, F, and G are snapshots at 50 minutes, 200 minutes, 350 minutes, and 400 minutes after the first cell division, respectively. F and G show the gradual bending of the embryo into the "comma" state.

The coloring of cells can be selected from two basic classifications: either by cell fate or by their mother cells (Table 1). Coloring of cells can be selected to suit the purpose of the visualization. Information is held in the cell color assignment map in the system. The default assignment map is based on cell fate, as shown in A. An alternative color map assigns colors to each cell based on the cell precursor (shown in B). All images shown in Figure 1 use the default color map. If 'transparent' is assigned to a cell, the cell will not be visible on the image, although it exists for computing dynamics. Using this feature, it is possible to display only a certain cell, making the other cells transparent. The system allows for zooming in and out, and rotation around the X-, Y-, or Z-axis. Simulation can be paused at any time step during the execution, allowing users to examine the visualized image as a still picture. Currently, all cells are shown as spheres. The use of more accurate cell shapes is possible, but at a much greater computing cost. Deformations in a cell's shape and the forces involved are approximated by overlapping spheres, and the repulsive force is proportional to the degree of overlap. A future system will implement cell shape deformation and an elastic model.

| Color      | Cell Fate   |
|------------|---|
| Red        | Dermal cells  |
| Green      | Neural cells  |
| Blue       | Digestive system cells                                  |
| Cyan       | Body muscle cells                                       |
| Magenta    | Germ cells  |
| Yellow     | Excretory system cells                                  |
| White      | Miscellaneous blast cells<br>(also mixed lineage cells) |
| Dark Green | Tail spike cells  |
| Dark Grey  | Cells which will die                                    |
| Black      | Dead cells<br>(not yet absorbed by other cells)         |

(A)

| Color   | Precursor Cell             |
|---------|----------------------------|
| Red     | AB                         |
| Green   | MS                         |
| Blue    | E                          |
| Cyan    | P,Z                        |
| Magenta | D                          |
| Yellow  | C                          |
| White   | Early generation (ex. Egg) |

(B)

Table 2: Color Maps

## 5 Neural Circuits

The neural system of *C. elegans* consists of 305 neurons with over 5,000 connections. The circuit topology is fully identified, although it may contain some errors, and various biological investigations are being made. In addition, a machine readable form of circuit topology data is avail-

able [Achacoso and Yamamoto, 92]. A few reports exist on the simulation of parts of the neural system of *C. elegans*, such as a mechanosensory subsystem [Wicks et al., 96, Wicks and Rankin, 95].

In conjunction with embryogenesis model, we are developing a simulator for neural systems in order to identify the functions of each component in the circuitry. Initially, we have built a simulator that focuses on a neural subsystem involved in thermotaxis. Thermotaxis is an example of *C. elegans*'s learning and memory behavior. When *C. elegans* is kept in a thermal gradient dish and given a food at a certain place, it memorizes the temperature of the area where it got the food. Later, it tends to move toward the temperature zone where it got food in the past. A neural subsystem for thermotaxis is mainly composed of 20 neurons (12 sensory neurons, seven inter-neurons, and a single motor neuron) [Mori and Ohshima, 95].

Our simulator can generate arbitrary input stimuli and simulate the possible voltage levels of each neuron. Information about even local neural dynamics in *C. elegans* is far from complete. We hope to improve the simulator so that it can incorporate new results from laser ablation experiments, mutant analysis, etc.

Figure 2 shows a screen of a simulation system for the neural subsystem responsible for thermotaxis. Arbitrary stimuli can be given in the form of an electrical voltage change on the sensor neurons. In this neural circuit, a continuous voltage change is propagated instead of spikes, which reflects the real physiological characteristics of the *C. elegans* neural system (*C. elegans* neurons are believed to have no action potential). The polarity, delay, and conductance of the connection can be changed with the simulator.

## 6 3D Visualization in Java

After the development of our 3D *C. elegans* model, microscopy data is now becoming available which, to a limited extent, gives accurate three-dimensional position information for many preembryonic cells [Martinelli et al., 97]. This data, along with gene sequencing, cell lineage data and known neural connections, is now freely available at the *C. elegans* database on-line at <http://probe.nalusda.gov:8300/cgi-bin/browse/acedb>. This data is still sparse, but we will soon incorporate it into our 3D C++ cell model to assist its model-based cell position estimation.

We have also created a 3D visualization system written in 100% Java which uses this new data to show the position and early migration of preembryonic *C. elegans* cells. Figure 3 shows a screenshot of this system. This system is different from our C++ model in that its intent is not to *simulate* cell dynamics, but only to visually display known biological data available on the web. The system not only displays 3D position data, but also neural synaptic connections, cell fates and expression patterns, lineages, etc. For completeness, the visualization system can also show postembryonic cells; since

postembryonic cell location and migration information is currently not available (and there is a *lot* of postembryonic cell migration, beginning with the “comma” migration), the system just assigns postembryonic cell positions near their parents' positions based on a simple cell-splitting lineage heuristic.

The visualization system comes with a number of options. The system allows the user to find cells by name, display chemical synapses and gap junctions, and highlight neurons, cells of a given cell fate, cells with a given expression pattern, or the descendants of any particular cell.

Because we have implemented the system in Java, it is highly portable. However, Java is new and necessitates some compromises. As efficient, portable 3D rendering technology for Java is currently nonexistent, the simplest version of our Java visualization system draws cells simply as points indicating the position of the center of the cell. Synapses are similarly displayed simply as lines connecting cells. A more advanced version of the visualizer, which runs on machines which support Sun's Java3D package, shows cells as true rendered spheres. Since known three-dimensional cell data is sparse (perhaps two data points per cell on average), the system must interpolate positions between known data timestamps. In contrast to our more sophisticated C++ visualization system, for efficiency's sake the Java system currently uses only simple linear interpolation.

Nonetheless, we have found the Java simulator to be a useful tool for visualizing known three-dimensional embryogenesis data, cell types, and especially neural interconnections in *C. elegans*, primarily because of the wide range of data it can display. When better data becomes available, we hope to augment this system with post-embryonic cell migrations and positioning, giving the user a full tour of *C. elegans* cell lineage from the single zygote to the final worm, complete with neural connections and other useful cell data.

## 7 What's Next: Cell Fate Determination

It was once believed that the cell fate of *C. elegans* is autonomously determined, without the interaction of other cells. However, recent findings suggest that extensive cell to cell interactions are necessary to determine cell fate. In the first phase of the Perfect *C. elegans* project, we intend to develop a simulation system which can replicate various biological experiments related to cell fate determination in the early stage of embryogenesis. We will particularly focus on cell fate determination until the 46th cell stage, which is about 2 hours after the first cell division. This is because there have been extensive investigations at this stage concerning cell to cell interactions, the genes involved, and their mutations. This stage also represents important aspects of cell fate determination mechanisms, which can be generalized to clarify the cell fate determination of other model animals.

A brief explanation to illustrate this process would aid in understanding this issue. When an egg divides, P granules lo-

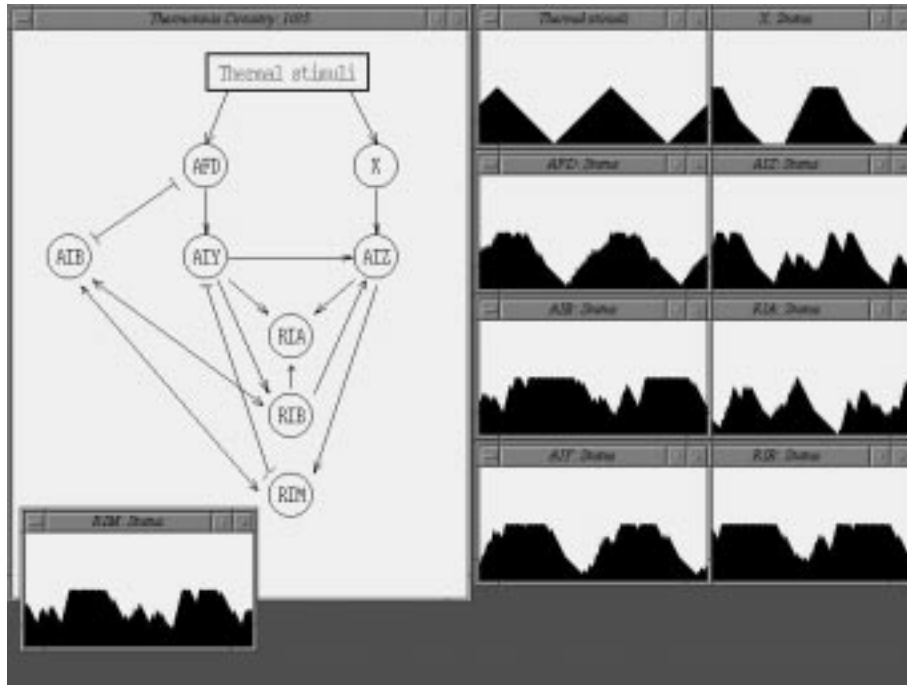


Figure 2: A Simulator for Thermotaxis subsystem

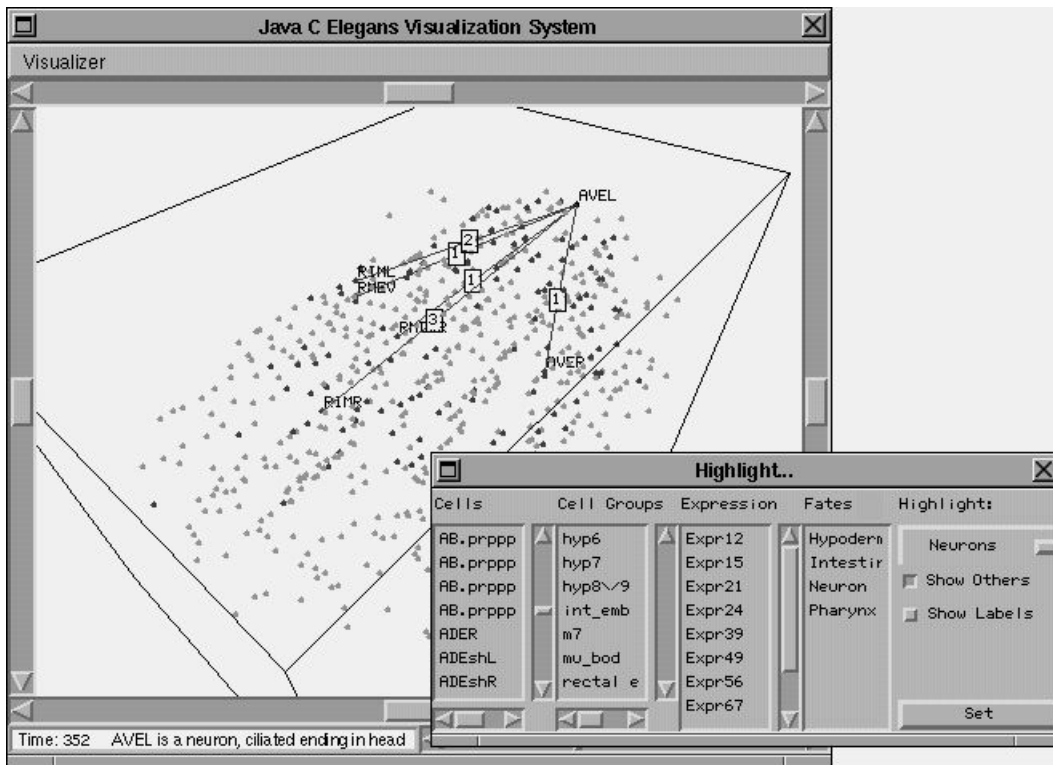


Figure 3: The Java 3D Visualization System for *C. elegans*



calized on one side of the cell; as a result, when the egg divides into the daughter cells P1 and AB, the P granules wind up primarily with the P1 cell. At the division of the P1 cell, P granules are again localized to one side to subsequently primarily appear in one of P1's daughters, P2. Hence, the localization of P granules determines cell fate autonomously. However, the determination of cell fate for the cells EMS and ABp, for example, are influenced by cell to cell communication. At the surface of the P2 cell, the genetic agent APX appears; through contact with the P2, the ABp and EMS cell membranes receive APX, which causes them to express the agents GLP-1 and mom-2, respectively. Should those cells be isolated from the P2 cell, these genetic agents are not expressed and cell fates of ABp and EMS will change. An overview of this process can be found in various sources [Priess and Thomson, 87]. Currently, we are developing a simulation model which enables the replication of such mechanisms.

Specifically, we will create a simulation system which can replicate known experiments on the effect of cell positions and interaction on axis determination. These experiments include loss-of-function mutant analysis for:

- the anterior-posterior axis determination (*par family*)
- cell fate determination of the founder EMS and  $P_2$  cells (with mutation of genes *skn-1* [Bowerman et al., 92, Bowerman et al., 93], *pie-1* [Mello et al., 92], and *mex-1*)
- cell fate determination for ABa and ABp, with dorsal-ventral axis and left-right axis determination (using mutants of genes *glp-1*, and *apx-1*, and laser ablation)

Other genes such as *emb-5* and *lin-12* will be modeled after establishing the simulation technique. Once we develop a simulation system which can accurately reproduce the above phenomena, it will be applicable to a wide range of biological systems. In order to make this part of the simulation as accurate as possible, we are planning to design a system to accept the National Institute for Genetics' *in situ* hybridization data [Tabara et al., 96].

## 8 Discussion

The ultimate goal of this work is to establish a powerful scheme for computer-aided biology, *bio-modeling*. Contrary to many artificial life research approaches, which start from a biologically-inspired but usually abstract dynamic model, our research is based on biological findings available at present and intends to look carefully into what actually takes place in life as we know it. By establishing a detailed model of a specific biological system, (in this paper, *C. elegans*), we intend to establish a more solid theoretical method for biology.

In modern particle physics, the work of theorists is essential. They formulate theories and predict hypothetical particles at precise energy levels. Experimental physicists then try to discover these predicted particles. For example, Glashow,

Weinberg, and Salam the *electro-weak theory* which explains the interaction of electromagnetic forces and weak forces. As an essential element of the theory, it predicted the existence of a set of hypothetical particles called "intermediate vector bosons". Carlo Rubia, an Italian experimental physicist, lead a team at CERN to discover this particle. The particle was discovered and the theory confirmed. This is a typical success story in particle physics.

We wish to bring such successes to biology, by providing a theoretical model which can not only display known data but make specific predictions which can then be verified or rejected through experimental effort. Due to the complexity of biological systems, it is natural that a computational approach be taken instead of the analytical approach of theoretical particle physics.

Because this project has just been started, the Perfect *C. elegans* Project is nowhere near the stage necessary to predict unknown and significant biological mechanisms. However, we are learning what is necessary to model biological systems to the level of detail needed to make meaningful predictions. Some of our findings are as follows:

- Choosing the right level of abstraction is essential. If we tried to build a simulation from the bottom up starting with protein dynamics, we would never reach a sophisticated enough model to simulate *C. elegans* embryogenesis. Not enough is known about molecular dynamics to simulate macroscopic phenomena such as embryogenesis. However, to totally ignore the molecular level is also not desirable, because any meaningful simulation must capture gene expressions and their interactions. We chose to simulate genetic interactions at the level of concentrations of promoters and repressors, and phenotypical effects are modeled functionally. We feel this is the right level of abstraction at this moment.
- Major parts of the simulation system can be applied in a general manner to simulate other animals simply by changing the underlying database. A general morphogenesis simulation system consists of (1) a kinetics simulation, which computes forces between cells, and (2) a genetic interaction simulation, which simulates the expression and repression of genes, the interaction of the products of genes with other genes, and cell-to-cell interactions.
- A study should be carried out to systematically search for all possible interactions of specific phenomena. When focusing on a specific phenomena, such as the cell fate determination of a specific lineage, there are fewer genes which might produce identical effects. A method needs to be developed to systematically generate a set of hypotheses and to define a critical experiment to identify which one of these hypotheses is correct.
- A well designed visualization, even with a crude simulation, can be a useful tool. Due to the complexity of the

phenomena, an intuitive understanding of the overall process of embryogenesis and neural dynamics is extremely difficult. Any system which assists in the understanding of complex phenomena would be welcomed by biologists, and is essential to help gather detailed data for more elaborate simulation systems. In this regard, the current version of the simulation system, particularly the visualization module, has been very successful. It generates reasonably accurate simulations, and visualization is very persuasive; many biologists working on *C. elegans* have welcomed our work and express interest in data exchange.

The future course of our research will therefore run along the following lines:

- Establishment of a simulation technique and the development of a general morphogenesis simulator.
- Establishment of a systematic method to generate possible genetic interactions and neural functions and to design critical experiments.
- Completion of the first version of the Perfect *C. elegans* system, which focuses on functions as a research tool for *C. elegans* biologists.

## 9 Conclusion

The work reported here is a first small step toward a synthetic approach to *C. elegans* biology. A series of simulation techniques have been developed for *C. elegans*, but which are applicable to the development of various organisms.

First and foremost, we have developed a useful, realistic simulation of *C. elegans* embryogenesis. The accuracy of the model can be improved with more data and by implementing detailed kinematic simulations. Simulation beyond 600 minutes can be made possible by adding this data. Although the current system is primarily for the visualization of the development process, it can be augmented to integrate genetic data such as a DNA sequence database, the pattern of gene expressions obtained by whole mount *in situ* hybridization, and other genetic information, to do modelling and prediction.

We have also created a Java-based visualization tool which displays known embryogenesis, neural interconnection, and other cell data. As available data grows, we think this will become a very useful tool for visualization. Further, we have implemented a simulator of the neural subsystem for thermotaxis, and are currently using it for analysis. Lastly, we are now implementing a genetic interaction simulation system, specifically targeting cell fate determination at and before the 46th-cell stage.

These initial simulations have produced promising results. While major efforts are required to achieve the level of detail and accuracy necessary for significant predictions to be made, the project has already identified some key issues in modeling

biological systems which can be generalized to other simulations. We believe that we have made a promising first step on what is undoubtedly a long journey.

## Acknowledgment

We would like to thank Jonathan Luntz, who did much of programming work for the SGI simulation system during his stay at Sony Computer Science Laboratory in the summer of 1996, supported by the NSF Summer Institute of the National Science Foundation. We would also like to thank Dr. Yuji Kohara at the National Institute for Genetics, and Dr. Richard Durbin at The Sanger Center for his valuable comments on the system.

## References

- [Achacoso and Yamamoto, 92] Achacoso, T., and Yamamoto, W. (1992) *AY's Neuroanatomy of C. elegans for computation*, Boca Raton, FL: CRC.
- [Bowerman et al., 93] Bowerman, B., Draper, B., Mello, C., and Priess, J., (1993) "The Maternal Gene *skn-1* Encodes a Protein That Is Distributed Unequally in Early *C. elegans* Embryos" *Cell*, 74, 443-452.
- [Bowerman et al., 92] Bowerman, B., Eaton, B., and Priess, J., (1992) "*skn-1*, a Maternally Expressed Gene Required to Specify the Fate of Ventral Blastomeres in the Early *C. elegans* Embryo" *Cell*, 68, 1061-1075.
- [Brenner, 63] Brenner, S. (1963). *A letter to Max Perutz*, 5, June. Downloaded from <http://eatworm.swmed.edu/Sydney.html>
- [Brenner, 74] Brenner, S. (1974). "Genetics of *Caenorhabditis elegans*," *Genetics*, 77, 71-94.
- [Sulston and Horvitz, 77] Sulston, J. E. & Horvitz, H. R. (1977). "Post-embryonic Cell Lineage of the Nematode, *Caenorhabditis elegans*" *Devl Biol.* 56, 110-156.
- [Kimble and Hirsh, 79] Kimble, J. E. & Hirsh, D. I. (1979) *Devl Biol.* 70, 396-417.
- [Martinelli et al., 97] Martinelli, D. (1997). "Gene expression and development databases for *C. elegans*", *Seminars in Cell and Development Biology*, 80, 450-467.
- [Mello et al., 92] Mello, C., Draper, B., Krause, M., Weintraub, H., and Priess, J. (1992) "The *pie-1* and *mex-1* Genes and Maternal Control of Blastomere Identify in Early *C. elegans* Embryos" *Cell*, 70, 163-176.
- [Mori and Ohshima, 95] Mori, I. and Ohshima, Y. (1995) "Neural regulation of thermotaxis in *Caenorhabditis elegans*", *Nature*, 376, 27.

- [Priess and Thomson, 87] Priess, J. and Thomson, J. N. (1987) "Cellular Interactions in Early *C. elegans* Embryos", *Cell*, 48, 241-250.
- [Sulston et al., 83] Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. (1983). "The Embryonic Cell Lineage of the Nematode *Caenorhabditis elegans*," *Dev. Biol.* **100**, 64-119.
- [White et al., 86] White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. (1986). "The Structure of the Nervous System of the Nematode *Caenorhabditis elegans*," *Phil. Trans. R. Soc.* **314**, 1-340.
- [Sulston et al., 92] Sulston, J. E. *et al.* (1992). "The *C. elegans* genome sequencing project: a beginning," *Nature* **356**, 37-41.
- [Tabara et al., 96] Tabara, H., Motohashi, T. & Kohara, Y. (1996). "A Multi-Well Version of *in situ* Hybridization on Whole Mount Embryos of *Caenorhabditis elegans*," *Nucleic Acids Res.*, **24** 2119-2124.
- [Waterston and Sulston, 95] Waterston, R. and Sulston, J. E. (1995). "The genome of *Caenorhabditis elegans*," *Proc Natl Acad Sci USA* 92:24, 10836-40.
- [Wicks et al., 96] Wicks, S., Roehrig, C., and Rankin, C. (1996). "A Dynamic Network Simulation of the Nematode Tap Withdrawal Circuit: Predictions Concerning Synaptic Function Using Behavioral Criteria," *J. Neurosci*, 16 (12): 4017-4031.
- [Wicks and Rankin, 95] Wicks, S. and Rankin, C. (1995). "Integration of Mechanosensory Stimuli in *Caenorhabditis elegans*," *J. Neurosci*, 15 (3): 2434-2444.