# The Virtual Biology Laboratories: A New Approach of Computational Biology


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

**Shugo Hamahashi**
Dept. of Electrical Engineering
Keio University
Yokohama 223, Japan
shugo@mt.cs.keio.ac.jp

**Jun Kitazawa**
School of Media and Governance
Keio University
Fujisawa 252, Japan
kitazawa@mag.keio.ac.jp

**Koji Takao**
School of Medicine
Keio University
35 Shinanomachi, Shinjuku
Tokyo 160, Japan
cozy@sun.microb.med.keio.ac.jp

**Shin-ichirou Imai**
School of Medicine
Keio University
35 Shinanomachi, Shinjuku
Tokyo 160, Japan
imai@sun.microb.med.keio.ac.jp



## Abstract

The Virtual Biology Laboratories is an attempt to create detailed biological models which can be used for research in molecular biology, developmental biology, and neuro-science. The goal is to create detailed models of specific model animals, so that various virtual biological experiments (such as gene knock out, microscopic manipulations, site direct mutagenesis) can be performed, and the results compared with actual biological experiments. Currently, we are working on three specific projects — Perfect C. elegans Project, Virtual Drosophila Project, and Virtual Cell Laboratory Project. These projects aim at detailed modeling of *C. elegans, Drosophila,* and the aging mechanism of human fibroblast, and Yeast *saccharomyces cerevisiae.* This paper describes the basic approach and some early results of the project.


## 1 Introduction

Modern biology as represented by molecular biology started when Watson and Click discovered DNA structure in 1953 [Watson and Click, 53]. Coincidently, the modern form of the computer, a stored program computer or the von Neumann architecture, was invented only a few years prior to the discovery of DNA. Perhaps, these are two most significant scientific breakthroughs in the mid-20th century. Computational biology, or computer aided biology, is where these two areas merge. Although there are substantial efforts to use computer systems in the area of genome data-base, motif extraction, and structure prediction, no systematic effort has been made to create a comprehensive model of an entire living system. We believe that it is a time to start an initiative to create complete detailed models of biologically significant model animals on computer. The goal of this paper is to provide a perspective on how computers can be used more actively in biology and in the way they will change how biology is performed. This paper describes what is envisioned, and discusses new ways in which biology is carried out.

### 1.1 *From Reductionist to Integrationist*

The fundermental scientific approach of the 20th century can be characterized as reductionist. Particle physics is a typical example. It made an ultimate challenge to subdivide this universe into subatomic particles as small as quark; and the space-time continuum and matter were reduced to super-strings. In spirit, the approach was applied to biology, which lead to the discovery of DNA as a basic information carrying component in the living system. Although there is no doubt that science clearly advanced by seeking fundamental components, it does not necessary enhance our understanding of this universe at a more macroscopic level. Identification of DNA does not mean that we understand the dynamics which give rise to the various phenomena inherent in life. Understanding of phenomena, which consists of many interacting components, is a important task. While most phenomena involve population dynamics and non-linear interactions, these phenomena often exhibit counter-intuitive behaviors. Even if the behavior of the system is not counter-intuitive, tracking the behavior of components in the system and emerged behavior of an entire system is almost impossible without the use of computers. This approach can be characterized as a synthetic understanding of biological system.

### *1.2 The Model of Scientific Success*

Given that the synthetic approach can be a powerful means to help understanding living systems, how should the computational approach be used in biology? Lessons can be learned from the major achievements in science. One of the most successful areas in modern natural science is particle physics. Although the ultimate theory – the theory of everything – is yet to be found, particle physics has discovered a set of basic principles which explain the fundamental mechanisms of matter and force. For example, one of the symbolic success stories involves the standard theory and the discovery of the intermediate vector boson. Glashow, Salam, and Weinberg proposed the electro-weak theory, which is now called "the standard theory", as a unified theory of electromagnetic and weak force. The theory predicted hypothetical particles called intermediate vector bosons (IVB), which had not been discovered at that time. Teams of experimental physicists at CERN tried to discover the IVB, and finally the UV-I team led by Carlo Rubia discovered the predicted particles called W and Z. By the discovery of these particles, the theory was confirmed, and the frontier of research was pushed to quantum color dynamics, which tries to explain the strong force responsible for interaction between quarks.

This story signifies how a new fundamental mechanism in nature can be discovered. In physics, particularly in modern particle physics, prediction using a promising theory has a decisive influence on the course of scientific inquiry. For a theory to be seriously considered, it must be able to explain a set of data which is already available. However, the truly powerful theory should be able to make significant and precise predictions which can be verified. The power of the theory lies in the prediction.

### *1.3 The Research Cycle of Computer Aided Biology*

How can we translate this model into the field of biology? Unlike particle physics, biology inherently deals with a large number of interacting complements. This essential nature of a biological system makes it difficult for the conventional mathematical approach to construct theories which are powerful enough to describe observed phenomena and to predict undiscovered mechanisms. While we do not deny the value of conventional mathematical analysis, we argue that for very complex and heterogeneous systems, the use of computer simulation should be considered as the central methodology.

Therefore, we propose that computer simulation which models mechanisms of biological processes should be used together with actual biological experiments, instead of using abstract mathematics describing average behavior of the system, so that results of simulation (which are virtual experiments) can be verified by tangible experi-

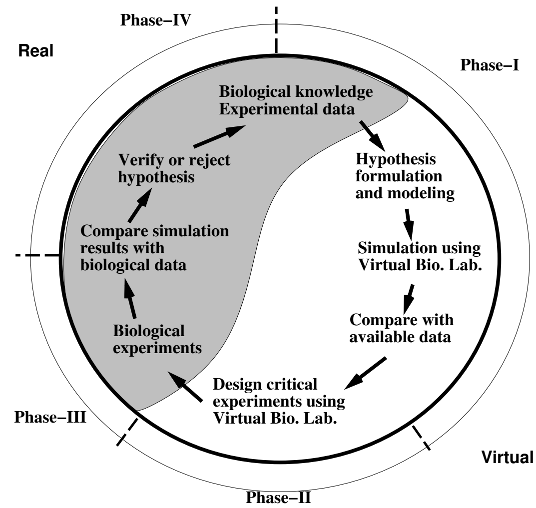



Figure 1: Computer Aided Biological Assay Cycle

ments. The actual procedure is illustrated in Figure 1. First, based on available biological knowledge and data, we create a detailed simulation model and carry out simulation. Once the simulation is performed, the results are compared with available biological data to check the validity of the simulation. Generally, the simulation results are compared with biological data from wild-type of the model animals. This is phase-I of the cycle.

Once the validity of the simulation is confirmed, another set of simulation will be done to identify a set of critical experiments which can provide us decisive evidence on specific biological mechanisms. In this process, several hypotheses on possible mechanisms will be implemented and compared with biological data. As a result, a set of critical experiments will be designed. This is the phase-II.

In Phase-III, actual biological experiments will be conducted based on the experimental design using the simulation system.

Finally, in the phase-IV, results of the experiments will be compared with the results of the simulation, so that the hypothesis can be verified or rejected. Whether the hypothesis is verified or rejected, it will add new knowledge on biological systems. A new cycle begins reflecting the results of the previous cycle.

The goal of this research program is to establish a new paradigm of biological research, and to promote understanding of living systems actually exist, rather than artificial life. It is important that we should be able to promote understanding of existing life, because it is the only instance of life we can agree on.

## 2 Overall Perspectives

Virtual Biology Laboratories is a research program consisting of several specific research projects under the unified goal of creating a new paradigm of biology using computers. In order to make a case for the proposed approach, and to make major breakthroughs, comprehensive research programs using selected high impact model animals and focusing on the modeling of biologically significant phenomena are needed. Therefore, we have started following projects;

**Perfect C. elegans project:** This project aims at detailed modeling of *Caenorhabditis elegans (C. elegans).* The advantage of using *C. elegans* for modeling is the abundance of available data. Complete cell lineage and neural circuit topology is known, and DNA sequencing is expect to be completed within a year or so. Expression patterns of all genes will be made available within a few years.

**Virtual Drosophila project:** *Drosophila melanogaster* is yet another extensively studied animal. Specifically, significant amount of data is available for early embryogenesis, such as gap genes, segment polarity genes, and homeotic genes. Various mutants have been identified, and detailed expression patterns have been studied. In addition, *Drosophila* has a much more complex morphogenesis than *C. elegans* and a substantial central nervous system. Formation of eyes and legs are also well studied. These conditions make *Drosophila* a most ideal subject for computational modeling.

**Virtual Cell Laboratory Project:** Virtual Cell Laboratory emphasizes simulation of the cell itself and *in vitro* cell culture. Specific targets are aging related mechanisms of human fibroblast, and Yeast *saccharomyces cerevisiae.* Cell aging was selected because it is one of the most significant issues in recent molecular cell biology, and it may involve heterochromatin and other novel regulatory mechanisms. Yeast is a natural choice as a model system for computer simulation, due to the abundance of available data. Mechanisms for cell cycle control are the most well investigated and significant parts of Yeast research.

In this paper, these projects are described using a set of initial accomplishments. While details of each project will be described elsewhere, it is important that the entire research program be described, even though rather sketchily, so that our intentions and approach can be made clear from various perspectives.

## 3 Perfect C. elegans Project

The Perfect C. elegans project aims at a detailed replication of *C. elegans.* When Sydney Brenner proposed a new project to investigate *C. elegans* to the Medical Research Council, he chose it because it is the simplest possible differentiated organism [Brenner, 63]. The decision was right, and a number of fruitful results were generated, including complete identification of the cell lineage [Sulston and Horvitz, 77; Kimble and Hirsh, 79; Sulston et al., 83], and neural circuits [White et al., 86]. The complete DNA sequencing and expression pattern mapping using whole mount *in situ* hybridization are now within our reach [Sulston et al., 92; Tabara et al., 96]. These investigations clarify what constitutes *C. elegans* and how each components operate. However, understanding the components and their isolated functions does not lead to an understanding of the dynamics behind the development and behavior of this organism. Despite the fact it is the simplest possible differentiated organism, it is still too complex to understand the dynamics and interactions taking place. It is obvious that a synthetic approach needs to be undertaken for better understanding of this intensively studied animal.

Thus, the goal of the Perfect C. elegans project is the creation of a detailed model of *C. elegans.* We intend to implement models for the embryogenesis process, genetic interactions, and neural circuits.

### *3.1 The Nematode Caenorhabditis elegans*

*C. elegans* is a small worm ubiquitously observed in the soil. It composes the largest biomass on earth. It has a life span of about 3 days and feeds on bacteria. The adult male *C. elegans* has 1031 somatic nuclei, while the adult hermaphrodite has 959 somatic nuclei. The cell lineage of these cells have been fully identified by the extensive work of a group of researchers, notably by [Sulston et al., 83].

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

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

The mechanism of fate determination involving maternal genes has been intensively investigated. However, subsequent fate determination is largely unknown because downstream genes have not been identified. In order to investigate genetic interactions for fate determination, a project to identify genes that are expressed in specific cell lineage was initiated [Tabara et al., 96]. Kohara's project at the National Institute for Genetics uses *in situ* hybridization on whole mount embryos to identify expression of genes at specific cells and specific times during the 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 genes cause cell lineage abnormality. There is a large list of genes and their phenotypical disorders. In addition, a series of manipulations are possible during embryogenesis through laser ablation or direct micro-manipulation.

### *3.2 Embryogenesis Visualization and Simulation*

Given these biological accomplishments and on-going efforts, the perfect C. elegans project aims at creating a detailed simulation model, one that promotes our understanding of life, particularly *C. elegans.*

As the first step toward this goal, we have developed a computer graphics system which visualizes the developmental process of *C. elegans.* The system is based on the cell lineage and cell location data investigated by Sulston et al., along with various other literature. The system generates computer graphics images from the division of the first cell to around 600 minutes after the first cell division.

#### *3.2.1 Data Resource*

It is a non-trivial task to create a reasonably accurate computer graphics image based on the available data because information necessary to create three dimensional models is missing. The following information is available:

- the complete cell lineage chart
- hand-drawn pictures in 2-1/2 dimension
  - 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
  - most cells at 430 minutes
- qualitative descriptions of the shape of embryo
- qualitative description of disparity in the size of divided cells
- general information on migration

However, much essential information is missing. For instance, the three-dimensional position of the cells is only given for a subset of cells at only a few time points. In addition, the lineage chart only indicates the approximate time of the divisions and the rough direction of division, such as anterior-posterior, dorsal-ventral, or left-right. For example, the cell lineage of two ring inter-neurons (ADAL and ADAR) are shown as `ADAL AB.plapaaaapp` and `ADAR AB.prapaaaapp`. This means that an `ADAL` cell was derived from the founder cell `AB`, and was created after a series of divisions in the order of posterior (`p`), left (`l`), anterior (`a`), posterior (`p`), four anterior (`aaaa`) divisions, and two posterior divisions (`pp`). `ADAR`, which is a symmet ric counter part has a similar lineage, but made a division to be a right cell after the posterior division. However, in reality, the direction of cell divisions are not necessarily aligned with these axis. A new technique must be develop to understand about the approximate direction of divisions.

By the same token, the locations of cells are drawn as a series of two-dimensional figures. While the approximate two-dimensional location of cells can be estimated from these figures, it only provides crude position for depth (or vertical position) of cells. The vertical position of cells are shown only by using circles with three levels of thickness. Again, a new simulation method needs to be developed to estimate 3D locations of cells.

Obviously, a straightforward way would be to collect three-dimensional position and shape data one step at a time, so that computer graphics could be created based on these data without problems. However, this approach is very labor intensive because each cells at every time step must be manually identified and the position data collected. In order to collect such data, researchers need a crude 3D visualization tool to assist in the identification of cells and their positions. In addition, such an approach cannot be applied to visualize the development of organisms having only a limited amount of position information. Therefore, we need to develop a system which can generate reasonably accurate 3D computer graphics based on limited data sets.

#### *3.2.2 Merging Simulation with Data*

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 in the observed data.

A set of data obtained from actual observation must be represented in machine readable form. Table 1 shows a part of a data record representing cell lineage and cell positions taken from the Sulston data. R is a record type (`f` and `p` mean a full record, `x` means a terminal record), and $d_1$ and $d_2$ are the names of daughter cells. When a '*' is attached in front of a symbol, such as '*a', it is concatenated to the name of the mother cell to create the name of the daughter cell. For example, `AB` cell'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 $d_1$ cell is shown as 'vol'. The time of division is indicated in 'time'. $x_1$, $y_1$, $z_1$, $r_1$ are the x, y, z positions and the radius of the cell at time point $t_1$,

| 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

respectively. A set of data can be repeated for different time points. Basically, this data-base contains most of the information provided in [Sulston et al., 83]. Even when we obtain more accurate data, we can continued to use this data record by making minor changes. One point which needs to be changed is the way the direction of cell division is specified. In the current data structure, it is represented as a three-dimension vector where each dimension has only three values (1, 0, and -1). This will be changed to represent a continuous number. The polar coordinate system will be used. Cell shape information is not included in the current version. This needs to be added in the future.

Second, various simulation techniques are used to fill in missing information, such as the location of cells not given in the data. Essentially, this part of the system computes the dynamics between cells, such as the force to push back colliding cells (equations not shown). However, if only dynamics 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, a force that a cell is supposed to generate for its movement is estimated using the inverse kinematic technique, and added to the cell's force vector.

The shape of an embryo changes after 200 minutes creating the so called "comma" and "2-fold" shape. While the driving force bending the embryo's shape is not identified biologically, computer simulation impose a top-down kinematics. At 200 minutes, a cylindrical coordinate system is imposed to describe cell locations. The cylindrical coordinate system itself slowly bend to agree with the observed shape, so that the cells change their absolute position according to the bending of the cylindrical coordinate. As a result, we have successfully created a computer graphics visualization of embryogenesis of *C. elegans* (fig. 2).

Coloring of cells can be selected to suit the purpose of the visualization. This information is held in the cell color assignment map in the system. The default assignment map is based on the 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 uses 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, making it possible to display only certain cells.

| 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 (also mixed lineage cells) |
| Dark Green | Tail spike cells |
| Dark Grey | Cells which will die |
| Black | Dead cells (not yet absorbed by other cells) |

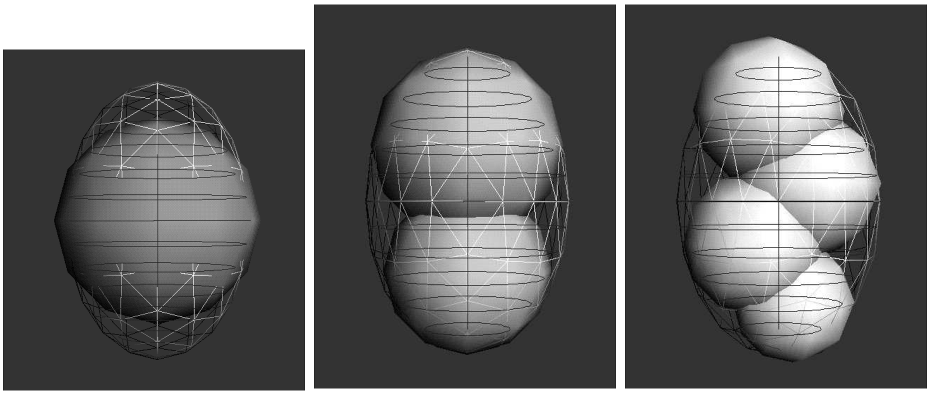
(A)

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

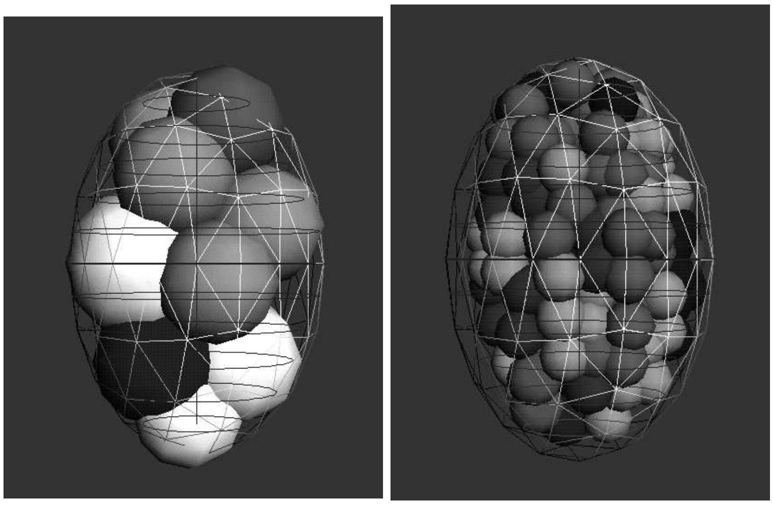
(B)

Table 2: Color Maps

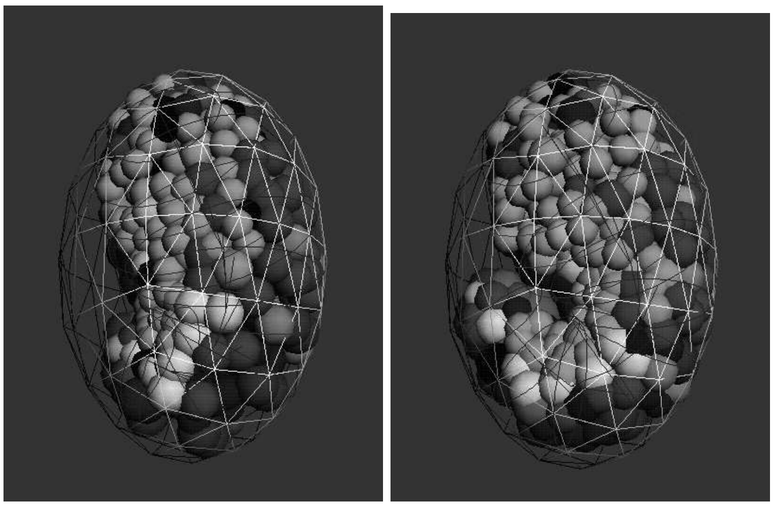
Current implementation uses C++ programming language, an OpenGL graphics library, and a SGI sphere library on a Silicon Graphics workstation. The cell coloring can be selected either by cell fate or by their mother cells (Table 1). The system enables zooming in and out, and rotation around any X-, Y-, Z-axis. Simulation can



(A) 1 cell (B) 2 cells (C) 4 cells



(D) 50 min. (E) 200 min.



(F) 350 min. (G) 400 min.

Figure 2: Snapshots of computer graphics images of C. elegans embryogenesis
Computer graphics (CG) images of the development of *C. elegans.* Colors assigned to cells based on the cell fate color map. A, B, and C are snapshots of the CG image at the initial stage (egg), two cells stage, and four cells 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.

be paused at any time step during the execution allowing users to closely examine the visualized image as in a still picture.

Implementation of the model can be an extremely useful tool for biologists working on *C. elegans*. It can be a comprehensive visual data-base of C. elegans, as well as being an assistant system for 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. Basically, it is 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. Our system, viewed as a visualization system, compliments the Angler system by providing computer graphics images and simulations linked with a set of optical images.

However, the system is not merely a visualization and data-base interface. It implements a physical model which also computes the dynamics between cells. Modeled as continuous elastic body, it can be augmented as a virtual laboratory in which we can conduct microscopic operations such as shifting cell location and laser ablation. We are currently working on the implementation of a genetic interaction simulator which can be integrated with the physical simulation system already developed. The new genetic components specifically model cell fate determination until the 46 cells stage, where major cell fate determinations are made.

In addition, we have already implemented a simulator for a thermotaxis neural subsystem [Mori and Ohshima, 95]. The implementation of simulator for a complete neural circuit is underway. Overall, the Perfect C. elegans project is at the phase-I level and is quickly moving toward phase-II.

## 4 Virtual Drosophila Project

The Virtual Drosophila projects aims at the detailed modeling of *Drosophila*, which is yet another extremely well-investigated model animal.

The embryogenesis of *Drosophila* differs from that of *C. elegans* in that it does not form cells at the earliest stages. After 13 successive nuclear divisions, about 6,000 nuclei form a syncytial blastoderm. Cellular membrane do not form until the thirteenth nuclear divisions. Thus, until this stage, gene products can spread freely in the embryo. The major part of the axis and body segment formation takes place during this period. This is a complementary subject to *C. elegans* where cell-cell interaction plays a major role in cell fate determination. By investigating these two model animals which have different embryogenetic mechanisms, we can better understand the essence of development.

As a first stage of the Virtual Drosophila project, we have implemented simulation for body axis and body segment determination. This is an early embryogenesis process, and involves maternal effect genes, gap genes, pair-rule genes, and segment polarity genes. Table 3 shows the set of genes which have been modeled at present.

| Class | Gene |
|---|---|
| Maternal Effect Genes | bicoid (bcd)<br>nanos (nos)<br>troso (tor) |
| Gap Genes | hunchback (hb)<br>Kruppel (Kr)<br>knirps (kni)<br>giant (gt)<br>tailless (tll)<br>hackebein (hkb) |
| Pair-Rule Genes | even-skipped (eve) |

Table 3: A list of genes implemented in a current version of simulator

The simulator should be able to model such processes as gene transcription, translation, and diffusion. First, gene transcription can be modeled by a stochastic process where activators and repressors compete for a binding site. When two components compete for one binding site, a simple way to approximate the probability on which a component binds to the site can be described as:

$$P = \frac{U_a}{U_a + U_b} \tag{1}$$

where $U_a$ and $U_b$ are the levels of protein $a$ and $b$ concentration near the binding site. However, this assumes that the binding affinity is equal, which is not necessary the case. Thus, by incorporating binding affinity, the above equation can be rewritten as:

$$P = \frac{\alpha_a U_a}{\alpha_a U_a + \alpha_b U_b} \tag{2}$$

where $\alpha_a$ and $\alpha_b$ are the binding affinity of the activator or the repressor $a$ and $b$, respectively. In a simple model, the amount of transcription directly corelates with the probability that the activator binds to the binding site. There are, however, much more complex situations, which would require more elaborate simulation. The model for such a case will be described elsewhere.

Transcription of a gene results in mRNA. In the simulation, mRNA is created as a product of gene transcription, and protein is created by translating mRNA. Some maternal effects gene products such as *nanos* binds to *hunchback* at translation process, not at the transcriptional process. Products of most other genes involved in this simulation can be assumed to act directly on the promotor region as an activator or as a repressor. The

| | bcd | nos | tor | hb | Kr | kni | gt | tll | hkb |
|---|---|---|---|---|---|---|---|---|---|
| bcd | | | | (+) | + | | | | |
| nos | | | | - | | | | | |
| tor | | | | | | | - | | |
| hb | | | | | + | - | - | | |
| Kr | | | | | | + | - | | |
| kni | | | | | - | | - | | |
| gt | | | | | - | - | | | |
| tll | | | | | - | - | - | | |
| hkb | | | | | - | - | - | | |

Table 4: Transcriptional Regulation Table

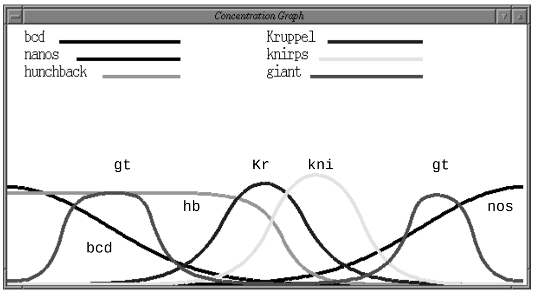



Figure 3: Simulated Expression Pattern of Gap Genes

regulatory relationship between genes and their products are shown in Table 4. In Table 4, + means that the protein acts as transcriptional activator of the target gene. (+) means that it promotes the translation process by acting on mRNA. − shows that it acts as a repressor. For example, *bcd* acts as a promotor of translation for *hb* and as a transcriptional activator for *Kr*.

Once the protein is created, it diffuses throughout the egg. As a crude approximation, we use a cylindrical coordinate system, instead of the exact oval shaped embryo. Currently, we are developing a more accurate 3D, oval egg so that a slight increase in concentration level at the anterior or posterior ends can be modeled. A simple equation for diffusion can be written as:

$$\frac{\partial U_i}{\partial t} = D_i \frac{\partial^2 U_i}{\partial x^2} \quad (3)$$

where $U_i$ is concentration of protein $i$, and $D_i$ is diffusion constant for $i$.

Figure 3 shows teh simulation result for expression patterns of gap genes for the wild type *Drosophila*. This is highly consistent with actual expression patterns of *Drosophila* wild types.

A pioneering study on simulation of *Drosophila* early embryognesis can be found in [Reinitz and Sharp, 95; Reinitz, et al., 95]. Our simulation system extends such a study by enabling creation of various mutants, such as the *giant* knock out, and the *bcd* knock out, which result in a loss-of-function type mutant, and the overexpresion

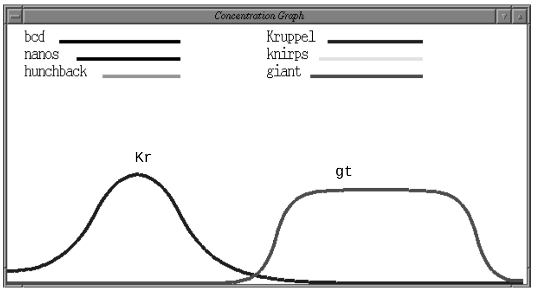



Figure 4: Simulated Expression Pattern of *bcd*- mutant

of specific genes. There are three major ways to carry out mutant analysis using the system.

The first method is the loss-of-function knock out, where a specific gene is knocked out so that no product can be created from that gene. Such experiments can be replicated by simply disabling the transcription of the target gene in the simulator. For example, Figure 4 shows a simulated gene expression pattern of the *bcd*-, *bicoid* knock out mutant. Only *Kr* and *gt* are expressed. Because no *bcd* exists, *hb* is not translated, and therefore is not expressed.

The second method is the overexpression of a gene, where the dosage of product from the target gene is artificially increased. This can be replicated by increasing a parameter which determines transcriptional efficiency, or by imposing a constant level of products to be present in the system.

The third method is site directed mutagenesis, in which a sequence of the specific binding site is altered disabling the activator or the repressor binding capability. For those promotor regions where a detailed model is implemented, this method can be reproduced by changing the binding affinity of specific products at a specific site.

Using these methods, the Virtual Drosophila system can replicate most of the mutation analysis actually carried out by molecular biologists. Thus, we can confirm the accuracy of the simulation. as it predicts possible molecular mechanisms of the embryogenesis process.

Currently, the Virtual Drosophila project is moving from Phase-II to the Phase-III level with regards to prediction on some specific gene regulation mechanisms, such as stripe 2 of *eve*.

# 5 Virtual Cell Laboratory Project

## *5.1 Overview*

Virtual Cell Laboratory (VCL) is a computer simulation system which implements abstract models of molecular dynamics on gene regulation for each cell. Contrary to previous projects which focused on *in vivo*, VCL targets *in vitro* biological experiments. For many biological systems which have complex population dynamics and nonlinear features, discreteness, and cell-cell interaction, attempts to draw a set of equations often fail to capture the essential dynamics of the system, and they often generate misleading data. Cellular senescence is one of many biological phenomenon, and embodies characteristics where average analysis face difficulties. Thus, our simulation system takes into account following issues:

**Cell-based model:** In our simulation, virtual cells are created, each of which represents an actual cell in a corresponding biological experiment. This feature enable us to precisely model behaviors of the heterogeneous cell population.

**Modeling mechanisms and interactions:** We simulate the internal mechanisms and the cell-cell interactions of each cell. Complex dynamics of histon complex, nucleosome, repressor, and cell-cell interactions are generally too complicated to describe using equations without losing essential characteristics. Instead of using equations, we create computer programs to simulate mechanisms inside the cells and their interactions. Each virtual cell is assigned an independent memory space in the main memory of a computer. Features such as, state of nucleosome binding sites, transcription, concentration of regulators and promoters are represented as a set of data structures and variables assigned to each cell. In the current, simulation we can create up to $10^5$ virtual cells.

**The level of abstraction:** We simulate abstract molecular mechanisms, rather than detailed dynamics of each molecule. For investigating phenomenon at the cell population level, modeling from molecular dynamics itself is too detailed and computationally expensive. Instead, we simulate the state of chromatin structure, such as the binding of nucleosome, or telomere shortening.

### *5.2 VCL Applied to Cellular Senescence*

As the first target of the VCL system, we have applied our approach to investigate the mechanisms of cellular senescence. Cultured diploid cells have a limited proliferative potential. This is known as the Hayflick limit [Hayflick and Moorhead, 1961]. This phenomenon has been confirmed by various experiments, and the age of the donor inversely corelates with the doubling potential of the fibroblasts. Numbers of genes associated with cellular senescence have been reported, and numbers of hypotheses have been proposed. However, none explain a comprehensive body of data which includes data that appear to be contradictory. Among many experimental observations, there are two major observations which a correct theory of cellular senescence must account for. These are growth kinetics data, proportion of non-dividing cells in the culture, and gene expression data. VCL was used to test various hypotheses with regards to cellular senescence. Figure 5 demonstrates the VCL screen applied to cellular senescence. It shows simulated data on growth kinetics, proportion of non-dividing cells in the culture, the physical location of cells with simulated reporter genes, and a simulated continuous northern blotting.

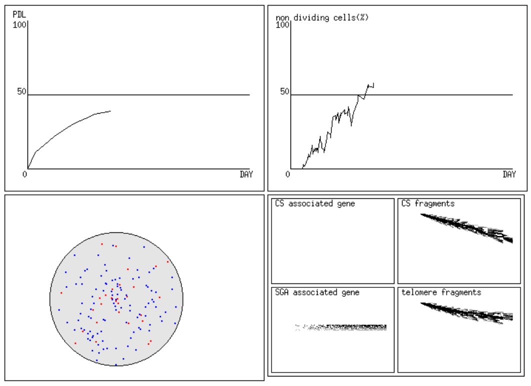



Figure 5: VCL Applied for Cell Aging Simulation

As a result of extensive simulations, we have come to the conclusion that heterochromatin structures control silencing of the genes related to the aging of the human fibroblast [Kitano and Imai, 1996]. The hypothesis not only explains all existing data on normal human fibroblasts, but also explains the data for Simian Virus 40 T Antigen-Transformed Fibroblasts (cells that are transformed by SV-40 T-antigen, which significantly extends their life-span), measured in terms of PDL, and Werner's Syndrome Fibroblasts. These fibroblasts taken from Werner's Syndrome (WS) patients, show significantly shorter life-span. In addition, predictions made by the simulation have been confirmed recently in actual cells [Imai et al., 1996].

The cellular senescence part of the VCL project have completed the first cycle of the computer-aided biological assay. Formulation of a more a detailed hypothesis will be stage-I of the second cycle. The yeast part of the project has just started and is still at the stage-I level.

## 6 Discussions

These projects are designed to complement the strengths and weaknesses of each project. These are inherent when choosing certain model animals. If, for example, *Drosophila* is chosen, it is not possible to use its cell lineage, because the number of cells is too large. *Drosophila*'s strength is in extensive analysis of the developmental process; also, their body structure is much more complex than *C. elegans*. *C. elegans* has a particular advantage in the existence of a complete cell lineage, DNA sequence and neural circuit data. However, its neural system and developmental process are too simple; therefore, knowledge obtained by using *C. elegans* cannot be directly applied to other, more complex animals. However, it is an extremely good subject for establishing methodologies and for understanding the basic behavior of developmental and neural systems. In addition, although major differences exist between *C. elegans,* humans, mice, and *Drosophila*, there are substantial commonalities. Extensive studies are now being made on the neurogenesis of central nerve system. VCL focuses on *in vitro* simulation, and tries to model more details of cell itself.

By combining the simulation skills developed for these model animals, we are able to design a generalized software system which can be used for a broad range of computer-aided biological studies. We are currently developing a physical simulation module, a genetic interaction module, and a neural system module. These modules are defined using a class library which defines teh behaviors of genes, cells, cell-cell interactions, and physical interactions. The level of details in the simulation can be adjusted, once the package is completed, by selecting the functions to be used from the class library.

## 7 Conclusion

In this paper, a trong argument for a new approach to be taken for more powerful usage of computer systems in computational biology, or computer-aided biology, is presented. The three projects described in this paper exemplify this approach. While it is not possible, due to space limitations, to explain the details of each project or to discuss significant predictions enabled by the approach, the activities of our biological counter-parts have been significantly reinforced and empowered. This is mainly because the modeling and simulations provide a clearer picture of possible hypotheses and scenarios which may actually be happening in the living systems. We hope that this paper contributes to the artificial life community in bridging the gap between artificial life and the natural life.

## References


[Achacoso and Yamamoto, 92] Achacoso, T., and Yamamoto, W. (1992). *AY's neuroanatomy of C. elegans for computation,* Boca Raton, FL: CRC.

[Brenner, 63] Brenner, S. (1963). *A letter to Max Perutz*, 5, June. Up loaded on `http://eatworm.swmed.edu/Sydney.html`

[Brenner, 74] Brenner, S. (1974). "Genetics of *Caenorhabditis elegans*," *Genetics*, **77**, 71-94.

[Hayflick and Moorhead, 1961] L. Hayflick and P. S. Moorhead. (1961). "The Serial Cultivation of Human Diploid Cell Strains," *Exp. Cell Res.* 25, 585.

[Kitano and Imai, 1997] Kitano, H., and Imai, S., "A Computer Simulation of Cellular Senescence," in preparation.

[Kitano and Imai, 1996] Kitano, H., and Imai, S., (1996). "The Virtual Cell Laboratory Predicts The Molecular Mechanism of Cellular Senescence — A Computer-Aided Approach to Biology of Aging," *Molecular Biology of the Cell,* Vol. 7, Supplement, 533a.

[Imai et al., 1996] Imai, S., et al. (1996). "OCT-1 Mediated Transcriptional Repressive Machinery Controls Cellular Senescence and Immortalization Associated Gene Transcription," *Molecular Biology of the Cell,* Vol. 7, Supplement, 533a.

[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). "The Post-Embryonic Cell Lineage of The Hermaphrodite and Male Gonad of *Caenorhadbitis elegans*," *Devl Biol.* **70**, 396-417.

[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.

[Reinitz and Sharp, 95] Reinitz, J. and Sharp, D. (1995). "Mechanism of *eve* stripe formation", *Mech. Dev.,* 49, 133-158.

[Reinitz, et al., 95] Reinitz, J., Mjolsness, and Sharp, D. (1995). "Model for Cooperative Control of Positional Information in *Drosophila* by Bicoid and Maternal Hunchback", *J. Exp. Zool.,* 271: 47-56.

[Sulston et al., 83] Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. (1983). "The Embryonic Cell Lineage of the Nematode *Caenohabditis elegans,*" *Dev. Biol.* **100**, 64-119.

[Sulston et al., 92] Sulston, J. E. *et al.* (1992). "The C. elegans genome sequencing project: a begining," *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.

[Watson and Click, 53] Watson, J. D., and Click, F. H. (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature,* 171: 737-738.

[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.