GEISHA SYSTEM:
An Environment for Simulating Protein Interaction

Masanori Arita Masami Hagiya Tomoki Shiratori
{arita, hagiya, blacky}@is.s.u-tokyo.ac.jp

Department of Information Science, Graduate School of Science,
University of Tokyo
7-3-1 Hongo Bunkyo-ku Tokyo 113 Japan

Abstract

Biological analysis of Drosophila embryogenesis has provided a model of protein interaction in segment formation. In this paper we introduce GEISHA system, which verifies and revises the rules of pattern formation in embryogenesis. The system consists of three parts: rule-based simulator, evaluator, and user interface. The simulator tests all the possible rule patterns, and the evaluator qualitatively evaluates results of the simulator; it searches for the desired pattern of protein expression. The user interface enables us to input or save data using GUI.

1 Introduction

Many biological systems and their functions came known in the recent prosperity of molecular biology. One of the widely believed concepts is that of a genetic switch, the function of DNA-binding protein, which turns on or off its function; this protein binds to a specific DNA-site, and promotes or represses the transcription of other proteins. The network of DNA-binding proteins is found in many parts of biological systems.

The segment formation in the embryo of Drosophila melanogaster (D. melanogaster), a kind of fruitfly, embodies a good example of this interaction of DNA-binding proteins which are translated from segmentation genes. These segmentation proteins specify each segment, which later evolves into a designated part of an imago (adult). So far, biologists found more than 30 proteins relating this segmentation, and now they establish the rough model of interaction among proteins. Many factors relating segmentation, however, are still unknown.

- Does a DNA-binding protein work really as a switch?
- Are there any other segmentation genes?
- Are there any other model to explain the segmentation?
- Can segmentation of other insects be explained using the current model?

Computer simulation can help answer these questions and guide the direction of biological experiments.

Here, we have to make the meaning of simulation clear. The word "simulation" means a model which imitates and predicts the behaviour of the real world. Simulation is, however, no exact replica. For a simulation, we extract some distinctive features from the real world according to our purpose, and then, we expect its result to suggest new knowledge such as a natural law governing the real world, or prediction of experiments. Therefore,

有田 正規, 萩谷 昌己, 白取 知樹: 東京大学大学院理学系研究科情報科学専攻, 〒113 文京区本郷 7-3-1

it is important to keep the link between the real world and a simulation. Another focus of interest is the potential of a model. We have to investigate what we can show and derive with a model.

Several biological simulations have been reported so far. Some systems [8, 4, 3] perform rule-based simulation focused on systems' qualitative features. These simulation rules are biologists' knowledge derived from numerous experiments. This means that these systems simulate no more than biologists' conceptual model derived from experimental results. Systems cannot help revise the present model, although biologists admit that additional experiments may correct or refine the current model.

Another simulation [7] using differential equations focuses on system's quantitative features. A differential equation model, however, is difficult to determine its parameters, especially when some of them are unknown. With such unknown parameters, the model may end up in an imaginary model.

In short, former systems only trace the biologists' conceptual view or newly produce an imaginary one. They do not contribute to the revision of biologists' knowledge. Simulation, however, is expected to give an insight into the real world. As for D. melanogaster, we expect a simulation which

- checks whether protein works really as a switch.
- determines whether present knowledge is enough for the explanation of segmentation.
- finds other models producing the same result.

We developed GEISHA (Genomic Environment for Interaction Simulation and Hypothesis Approximation) system, which optimizes the rules for proteins in segmentation. This system is totally different from the quantitative system using differential equations or from qualitative system using rule-based expert system. GEISHA treats both quantitative and qualitative model of protein interaction and aims to reduce the model from a quantitative one to a qualitative one. We give the system rules with only two relations among proteins: promotion and repression, and the system searches optimal threshold-values for the given rules to explain protein interaction in segmentation. If threshold-values are found and small changes in those values do not affect protein expression, we can conclude that actual embryogenesis are robustly regulated. If no rule is found, on the other hand, our assumption should be wrong. There may be some unknown proteins necessary for segmentation, the relations among proteins may be wrong, or segmentation may depend on finer quantitative regulation.

2 Fly Embryo

Our simulation target is the segment formation of D. melanogaster, a kind of fruitfly. GEISHA simulates the middle part (15%∼70%) of an embryo. Segmentation of D. melanogaster is investigated in detail, and biologists establish a hierarchical model [6, 2, 5] of protein interaction, though hypothetical.

First, we shall briefly explain the segmentation of D. melanogaster. A fruitfly breeds fast and spends as an embryo the first 16 hours of its life cycle, 10–14 days. The entire embryo period is divided into 16 stages, and Stages 4 and 5 are called segment formation stages. During these formation stages, about 5000 cleavaged yolks migrate out to the surface of an egg, and form a superficial monolayer. This state is called the syncytial blastoderm. Then membranes fall among the yolks to partition them into cells. This state is called the cellular blastoderm. During this cellularization, a segment, each with a defined future role, is regulated by the pattern of segmentation proteins comprising four groups: *maternal-effect, gap, pair-rule, segment-polarity.* These groups form a hierarchy of control relationship in this order.

2.1 Maternal-effect Genes

Segmentation begins with asymmetry of four groups of maternal-effect genes. One of them, anterior-posterior polarity, arises from a localized deposit of mRNA. The mRNA of *bicoid (*abbreviated as *bic)* gene diffuses from anterior to posterior, and with two other gradients, *nanos (nos)* and *torso*, partitions the blastoderm sideways into four parts. Two out of these three proteins, *bicoid* and *nanos*, seem to relate the segmentation of the middle embryo.

2.2 Gap Genes

There are 6 gap genes: *hunchback (hb), Krüppel (Kr), knirps (kni), giant (gt), tailless (tll),* and *huckebein (hkb).* The gradients of these proteins overlap to one another with different peaks, and further divide the blastoderm. These gradients are regulated by gradients of maternal-effect genes. Only 4 genes, *hb, Kr, kni, and gt,* seem to relate the segmentation of the middle part.

2.3 Pair-rule Genes

Pair-rule proteins subdivide the middle part of a blastoderm into 14 segments according to the expression of both maternal-effect and gap proteins. This 14 segments will form the entire body of an imago. There are 8 pair-rule genes, and each of these forms 7 stripes of 4-cell-wide at 4-cell-wide intervals. Some proteins express themselves mutual-exclusively.

2.4 Segment-polarity Genes

At least 10 segment-polarity genes label the subdivisions of each segment. Pair-rule proteins regulate these genes. For example, the gene *engrailed* appears in a series of 14 bands, each of 1-cell-wide, corresponding to the anterior part of each segment.

3 GEISHA System

GEISHA system is a rule-based simulator which simulates segmentation, a transition period from a syncytial blastoderm to a cellular blastoderm. The function of a protein as a switch is described in terms of $if \cdots then$ rule. Of these rules, two relations of proteins, repression and promotion, are fixed in the course of simulation, and the system searches for the optimal threshold-values for the relation.

The embryo model comprises a list of columns, and each column represents the ring-shaped band of cells, that is, horizontal slice of an embryo. Proteins produced in a band of cells are stored in the corresponding column. At each time-round in each column, the simulator

1. deletes old proteins.
2. produces new proteins.
3. diffuses proteins.

The simulator repeats this cycle for a fixed time-round. For each possible pattern of rules, the simulator produces the protein distribution.

The evaluator qualitatively evaluates all protein distribution and checks which rule gives the optimal distribution.

3.1 Protein Abstraction and Rules

GEISHA treats interaction among proteins. In reality, however, there are many other actors on a stage. A ribosome translates mRNA into protein before the mRNA collapses, and a protein promotes or represses the transcription of primary mRNA of other proteins.

$$\text{protein} \xrightarrow{bind} \text{DNA} \xrightarrow{transcript} \text{primary mRNA} \xrightarrow{process} \text{mRNA} \xrightarrow{translate} \text{protein}$$

This process, however, can be summarized into two simple protein interactions: “repress” and “promote”.

$$\text{protein} \xrightarrow{promote,repress} \text{protein}$$

The amount of protein is expressed in `real`, and the combination of amounts triggers promotion or repression of relating proteins. For example, if *bic* protein promotes and *nos* protein represses the translation of *hb* protein, the rule for *hb* is as follows.

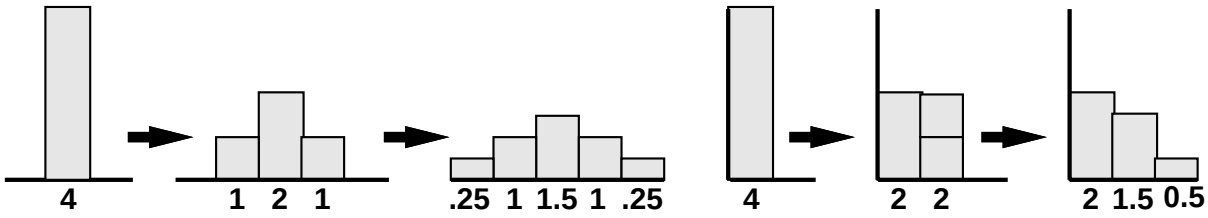

Figure 1: Diffusion

if $bic > 3.0$ & $nos < 1.0$ **then** $create(hb)$

Here, 3.0 and 1.0 are threshold-values for *hb*. We restrict threshold-values to `int`. Each protein has its own production rules of the above form; **if** part is a conjunction of conditions. Simulation rule is a set of rules of this pattern. You can see the entire rule in Appendix.

The simulator creates one unit of protein for each application of a rule. Each unit represents a certain population of molecules. This unit has a lifetime, which means the length of time during which the protein exists.

Since only one unit of protein is produced at one time-round, the maximum amount of protein is determined by this lifetime. For example, the lifetime of gap proteins are defined as 5, so the maximum amount of gap proteins is 5 units. Of course, actual half-life varys among proteins, but we assume that all proteins have the same lifetime, 5. Though this assumption makes the model less realistic, it reduces the complexity of the simulation.

3.2 Diffusion

In syncytial blastoderm, protein diffuses into neighboring cells, for membranes are incomplete to seal each cell. This diffusion effect is remarkable in the earlier stage; *bic* mRNA localizes at the anterior end, yet *bic* protein diffuses down to the middle of an embryo. In contrast, diffusion is not likely to occur in cellular blastoderm. For example, the pair-rule expression appears in gradients at first, but as time passes, sharpens its intensity and delineates the anterior boundaries of the stripes.

To simulate this effect, the simulator employs diffusion-cycle. In every time-round, each protein undergoes defined times of diffusion-cycle. In each diffusion-cycle in each column, $\frac{1}{4}$ of the amount of protein diffuses to both sides of the column. (Note that the amount is represented as `real`.) If one side is blocked, half of the amount of protein diffuses to the other open side (Figure 1). For example, *bic* should undergo over 200 diffusion-cycle to model its slope from the anterior end to the middle, while *hb* needs less to model its sharp fall like a cliff.

3.3 Evaluation

The simulation for a rule with given threshold-values halts after 20 time-rounds. Then, the expression of proteins is transformed into intervals, and the difference between the optimal result and the simulated one is compared by the evaluator. Optimal intervals of proteins contains "don't cares"(Figure 2, right below), and this pattern is given by a user. If both results are the same, the evaluator preserves the result as a candidate, and makes the simulator run again with slightly different rules. The simulator exhaustively searches all patterns of threshold-values of the rule. The translation from protein expression to intervals proceeds as follows.

1. Divide the amount of protein into binaries at the threshold of 3.0.
2. Simplify the interval relation(Figure 2, above).
3. Match the common pattern of protein expression between the optimal and the simulated(Figure 2, right).

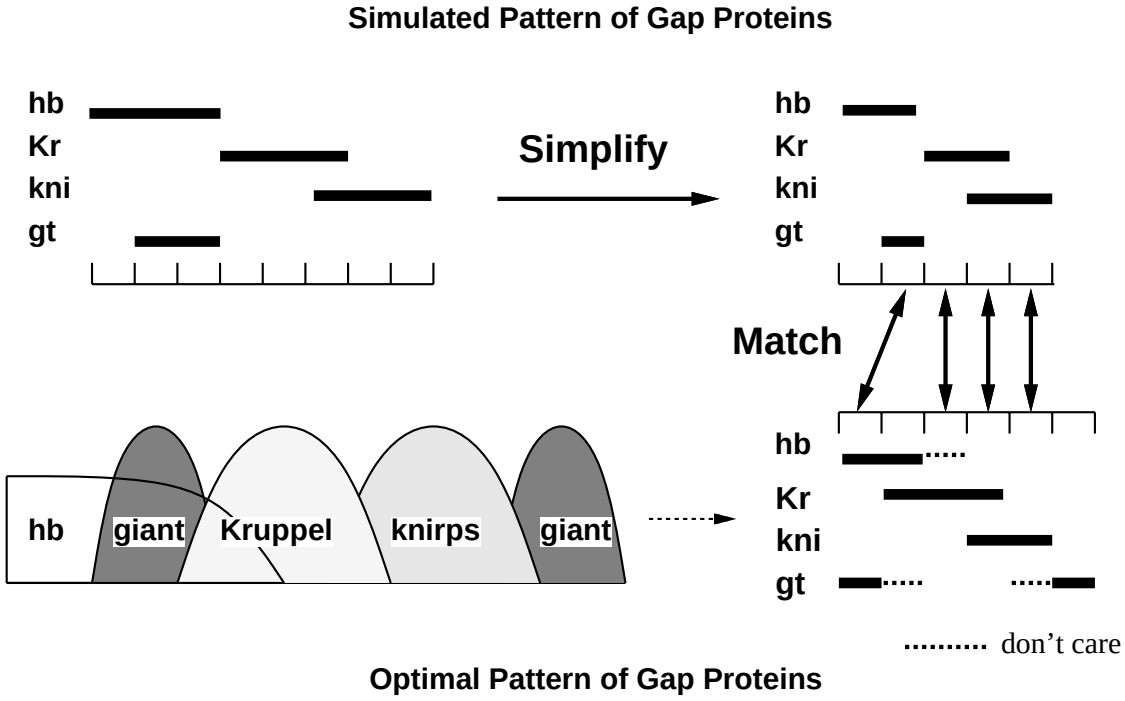

Figure 2: Convertion to Intervals and Match Process

The match procedure corresponds to a longest common subsequence problem, and its complexity is $\mathcal{O}(mn)$ (m and n are the length of two sequences.)

Figure 2 is the interpretation of the table below. This is a sample model, and each column corresponds to a band of cells. We interpret 3 or more amount of protein as 1, and 2 or less as 0.

hb	**5.0**	**4.5**	**4.0**	1.5	0.8	0.0	0.0	0.0
Kr	0.0	0.3	1.2	**3.5**	**4.2**	**4.2**	1.1	0.2
kni	0.0	0.0	0.0	0.0	0.0	**3.0**	**3.7**	**4.3**
gt	1.9	**3.1**	**3.5**	1.0	0.0	0.0	0.0	0.0

3.4 Simulation and Rule Modification

The exhaustive search of threshold-values soon bumps into the wall of combinatorial explosion. Therefore, we restrict threshold-values to 1, 3, or 5. (Note that threshold-values are `int`, and the maximum amount of protein is 5.) If the interaction among proteins were roughly regulated, we would find results similar to the optimal pattern without the finer tuning from the above three thresholds to other thresholds: 2 or 4. This assumption turned out to be true.

We tested GEISHA system with fixed gradients of maternal-effect proteins (*bic* and *nos*). The first goal is to find three rules for the gap-protein gradients (Figure 2, below). There are in total 12 threshold values in the three rules, but we fixed 4 of them. Search for $3^8 = 6,561$ patterns returns as many as 24 cases satisfying the desired result. This search took Sparc-station IPX about 67 min. (CPUtime). All the optimal thresholds is shown in Appendix.

target proteins	thresholds		pattern		CPU time (SPARC IPX)	
	total	searched	searched	optimal	user	system
Kür, kni, gt	12	8	6,561	24	3898.230	15.690
Kür, kni, gt	12	10	59,049	38	35136.750	149.370

3.5 GUI

User interface is an important factor of a simulation. We want to make this system available to biologists, therefore input and output of the system should be easy to examine for biologists. For this purpose, we built a user interface in X-window system (Figure 4). This interface provides us the following operations.

- Input, revise, or save rules.
- Input, revise, or save the protein expression.
- Show animation of the simulation.
- Examine simulation results interactively.

4 Discussion

4.1 Robustness

The simulator found lots of optimal threshold-patterns though we restricted the threshold-values to 1, 3, or 5. This means that a small change in a threshold-value, or equally, a small change in a concentration of proteins, does not affect the expression pattern of proteins. For example, a change in a threshold-value from 1 to 3, or from 5 to 3 does not affect the pattern in most cases. This observation confirms a multiple-gradient model rather than a single one (Figure 3). In a single-gradient model, a gradual slope of amount of one protein provides all information of locations, while in a multiple-gradient model, several gradients indicate information of locations. In the latter model, small changes in a gradient does not affect expression pattern, and this robust indication of locations confirms genetic switch hypothesis.

4.2 Redundancy

Many biological systems contain redundancy. For example, an embryo lacking both *nos* and *hb* proteins will normally grow up to an imago ([6] pp. 37). We cannot tell, however, which part of rules is redundant by seeing the set of threshold-values which produces optimal result. For example, we can remove the relation with *gap* protein from the rules for *Kür* and *kni* proteins (see Appendix). These threshold-values, however, cannot take all the possible values. A change of these threshold-values from 3 (or 5) to 1 results in a different pattern of protein expression.

We found many sets of appropriate threshold-values, but we could not find a set of rules with only two threshold-values: 3 and 5. In order to differ the *kni* and *Kür* expressions of we have to introduce at least three threshold-values.

Our set of rules is based on the result of biologists' experiments[1, 11]. Our *Kür, kni,* and *gt* rules contain no promoter, because there is no conclusive evidence that a promoter for these proteins exists. Biologists believe, however, that every protein should have at least one promoter, and the promoter for these three proteins is considered to be *bic*.

Our simulation shows an interesting result. As far as the formation of gap protein patterns, no promoter is necessary. We never think, however, there is no promoter. Both repression and promotion must be collaborating to form the correct pattern and to make the system robust.

5 Future Work

As is mentioned in the introduction, the simulator should indicate new knowledge about segmentation.

5.1 Redundancy Detection

It is difficult to tell whether there is redundancy in a given set of rules. We can check which part of rules is redundant just by deleting the part, but we have to find a good search strategy to predict which part may be deleted.

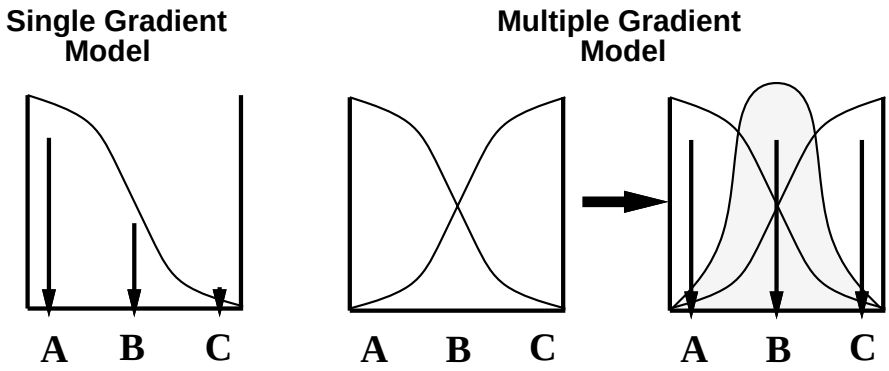

Figure 3: Gradient Model

5.2 Model Suggestion

We can change the relation among proteins, promotion and repression, and make the simulator run. By an exhaustive traverse of different thresholds, we may find sets of rules for explaining the same protein expression. If these sets of rules are found, they suggest that several models of segmentation may exist. On the contrary, we may find no rule. In this case, we can confirm the uniqueness of the relation which simulates the correct pattern.

Many mutation patterns are reported so far. If our model could explain these mutations, we could further confirm the correctness of our model. The difficulty lies in that much of the known mutations are mutations in pupa. We can never know how the expression of gap proteins is in an embryo.

5.3 Other Species

Biologists have investigated protein expression of some species other than D. melanogaster [9, 10]. In Callosobruchus, a kind of beetle, segmentation is similar to Drosophila, but in Schistocerca, a kind of grasshopper, segmentation shows a totally different pattern. The simulator can suggest rules for these other protein expression by traversing various rules.

6 Conclusion

We implemented GEISHA system for searching the optimal rules of pattern formation in embryogenesis of Drosophila melanogaster. This system simulates the protein interaction in the middle of an embryo, and showed that

- the hypothesis of genetic switch is credible.
- much redundancy exists in the model.

We want to further search for the optimal protein distribution with other rules, and build models for other species such as Callosobruchus or Schistocerca.

Acknowledgement

We owe a lot to Prof. Nakai and Prof. Suzuki at National Institute for Basic Biology, and Dr. Doi at Fujitsu Laboratories. Without their advice, this project was hard to proceed. We also thank Prof. Takagi at Human Genome Center for his kind criticism. In the course of making this system, he let us present our occasional report in his laboratory.

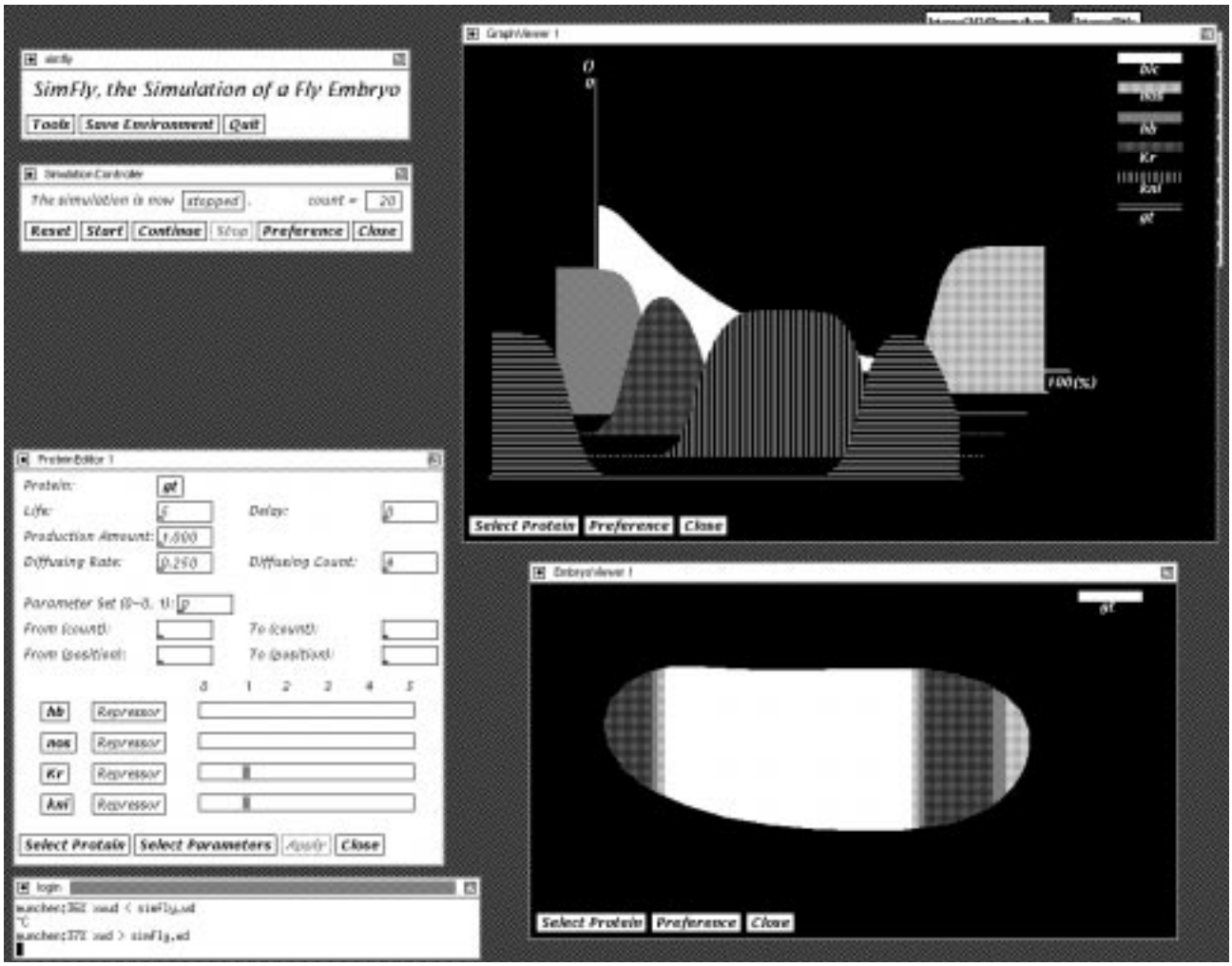

Figure 4: GUI

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from the Ministry of Education, Science and Culture of Japan.

References

[1] T. Gutjahr, E. Frei, and M. Noll. Complex regulation of early paired expression: initial activation by gap genes and pattern modulation by pair-rule genes. *Development*, 117:609–623, 1993.

[2] P.W. Ingham. The molecular genetics of embryonic pattern formation in drosophila. *Nature*, 335.1:25–34, 1988.

[3] P.D. Karp. Artificial intelligence methods for theory representation and hypothesis formation. *Computer Applications in the Biosciences*, 7(3):301–308, 1991.

[4] K. Koile and G.C. Overton. A qualitative model for gene expression. In *Proceedings of the 1989 Summer Computer Simulation Conference*, pages 415–421, 1989.

[5] A. Kuroiwa. *Homeobox(in Japanese).* Kodan-sya Scientific, 1989.

[6] P.A. Lawrence. *The Making of a Fly.* Blackwell Scientific Pub., 1992.

[7] H. Meinhardt. Models for maternally supplied positional information and the activation fo segmentation genes in drosophila. *Development*, 104:95–110, 1988.

[8] S. Meyers and P. Friedland. Knowledge-based simulation of genetic regulation in bacteriophage lambda. *Nucletic Acids Research*, pages 1–9, 1980.

[9] A. Michael. Is pairing the rule? *Nature*, 367:429–434, 1994.

[10] R. Sommer and D. Tautz. Segmentation gene expression in the housefly musca domestica. *Development*, 113:419–430, 1991.

[11] G. Struhl, P. Johnston, and P.A. Lawrence. Control of drosophila body pattern by the hunchback morphogen gradient. *Cell*, 69:237–249, 1992.

Appendix

24 optimal rules found by the system

3 1 1 3 5 5 1 1	1 3 1 3 5 5 1 1	1 5 1 3 5 5 1 1	3 3 3 3 5 5 1 1
3 5 3 3 5 5 1 1	3 1 1 5 5 5 1 1	1 3 1 5 5 5 1 1	1 5 1 5 5 5 1 1
3 3 3 5 5 5 1 1	3 5 3 5 5 5 1 1	1 5 1 3 5 5 3 1	3 3 3 3 5 5 3 1
3 5 3 3 5 5 3 1	1 5 1 5 5 5 3 1	3 3 3 5 5 5 3 1	3 5 3 5 5 5 3 1
3 3 3 3 5 5 1 3	3 5 3 3 5 5 1 3	1 3 1 5 5 5 1 3	1 5 1 5 5 5 1 3
3 3 3 5 5 5 1 3	3 5 3 5 5 5 1 3	1 5 5 1 5 5 3 3	3 5 3 5 5 5 3 3

Example.

"1 3 1 3 5 5 1 1" means

Kür		kni		gt			
kni	gt	Kür	gt	hb	nos	Kür	kni
1	3	1	3	5	5	1	1

Additional 14 optimal rules found by the system

1 3 1 1 1 3 5 5 1 1	3 3 1 1 1 3 5 5 1 1	1 3 1 3 1 3 5 5 1 1	5 1 5 3 5 1 5 5 3 3
1 3 1 1 1 5 5 5 1 1	3 3 1 1 1 5 5 5 1 1	1 3 1 3 1 5 5 5 1 1	1 1 5 1 5 1 5 5 3 3
1 3 1 1 1 3 5 5 3 1	1 3 1 3 1 3 5 5 3 1	1 3 1 3 1 5 5 5 3 1	
3 1 5 1 5 1 5 5 3 3	5 1 5 1 5 1 5 5 3 3	1 1 5 3 5 1 5 5 3 3	

Example.

"3 3 1 3 1 3 5 5 1 1" means

Kür			kni			gt			
hb	kni	gt	nos	Kür	gt	hb	nos	Kür	kni
3	3	1	3	1	3	5	5	1	1

$$\text{Kür} : \underbrace{(hb < 3)}_{fixed} \& \underbrace{(nos < 1)}_{always fixed} \&(kni < 1)\& \underbrace{(gt < 3)}_{removable}$$

$$\text{kni} : \underbrace{(hb < 1)}_{always fixed} \& \underbrace{(nos < 3)}_{fixed} \&(Kür < 1)\& \underbrace{(gt < 3)}_{removable}$$

$$\text{gt} : (hb < 5)\&(nos < 5)\&(Kür < 1)\&(kni < 1)$$

Rules for proteins

Name	Lifetime	Production	Diffuse		Production Location	Relations
			Rate	Cycle		
bic	5	5	0.25	256	Anterior End	$(bic < 1)$
nos	5	1	0.25	4	—	$(nos > 3)$
hb	5	1	0.25	4	—	$(nos < 1)\&(bic > 3)$
Kür	5	1	0.25	4	—	See above
kni	5	1	0.25	4	—	See above
gt	5	1	0.25	4	—	See above