SIMFLY2: Simulation of a Fly Embryo

Masanori Arita
arita@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 segment formation in Drosophila embryogenesis provides a good ground for modelling interaction of DNA-binding proteins. In this paper, we propose threshold model for qualitative simulation of the interaction, and introduce SIMFLY2. *This revised version of* SIMFLY, *a simulator for protein interaction, integrates genetic algorithm for the search of optimal relations among proteins. We confirmed that* SIMFLY2 *did find the relation we had found last time in* SIMFLY *by exhaustive search.* SIMFLY2 *also found interaction models between two pair–rule proteins and gap proteins.*

1 Introduction

Many DNA-binding protein molecules are said to have a symbolic function resembling a digital switch; they can turn on or off the activation of other protein molecules. In literature, the function of this genetic switch (GS) is characterized as activation, repression, or squelching [4]. This intuitive classification shows that GS works with crisp response, not with gradual or differential reaction. With this perspective exaggerated to the extreme, the function can be modelled with Boolean logic. However, quantitative features such as the concentration of regulatory molecules and their kinetic parameters let them act not the simple switch with `on` and `off`, but the multi-leveled volume switch. Then, what model instead of Boolean logic would be appropriate for this volume switch? So far, some enzyme reactions have been successfully described with differential equations. In most reactions, however, only qualitative relation is known and quantitative parameters are totally unknown. So we need more qualitative, intuitive model than the model with differential equations. Simple, yet eloquent model is the one we desire. In this paper we introduce "threshold model" for the qualitative simulation of this volume GS.

有田 正規: 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

The crispness of the GS, often understandably denoted using a sigmoidal curve, depends on the concentration of the proteins and the sensitivity of DNA sites. As the primary approximation instead of complex interaction at the molecular level, we consider only proteins, and no DNAs or RNAs. More specifically, only amounts of proteins are considered. Following is the basic interaction scheme in our model.

if `Threshold` α $>$ `amount(A)` $>$ `Threshold` β **then** `create(B)`

`A, B`	proteins
α, β	threshold-amounts of proteins

With these thresholds, we can describe following regulations.

Activation
if `amount(A)` $>$ `Threshold` β **then** `create(B)`
Repression
if `Threshold` α $>$ `amount(A)` **then** `create(B)`
Competition and Cooperation
if `amount(A)` $>$ ($<$) `Threshold` α **or / and**
`amount(B)` $>$ ($<$) `Threshold` β **then** `create(C)`

We applied our threshold model to an embryo of a fruit-fly. The embryonic segmentation of *Drosophila melanogaster* is famous for being regulated by complex genetic network (interaction among proteins), but its mechanism of interaction is little known, leaving most part to be elucidated in molecular biology. The prediction of this relationship would be a great help in determining the direction of biological experiments.

Last year we built SIMFLY, a simulator for checking and predicting relation among proteins, as well as for the grasp of expressive power of our threshold model. We simulated the relation among *gap* proteins using exhaustive search, and found the following result.

- The repressive power of *Kr* and *kni* proteins against *gt* are greater than that of *gt* protein against these.

This result assumes that these three *gap* proteins only repress with one another, an assumption which is hardly true in a real embryo. We had to consider other cases in which each protein has its activation site. We also wanted to further simulate *pair–rule* protein stage, formation of 7 stripes after *gap* protein stage. But, there was a problem: immensely vast search space. This time, we employed genetic algorithm (GA) for the search and found the seemingly good relation among *pair–rule* proteins and *gap* proteins.

In the next section how the embryo of *Drosophila* developes is briefly described. In Section 3, a related simulation of this embryogenesis is introduced and compared with our model. In section 4, our SIMFLY2 system – what threshold model is, how GA is used, and what result is obtained – is explained. The last section is allocated for discussion, current problems, and future work. After Conclusion, some rule patterns obtained by our system are described in Appendix.

2 Pair–rule Proteins

The process of segmentation in *Drosophila* blastoderm proceeds in a hierarchical manner in which the middle part of an embryo is stepwise divided into 14 segments. Segmentation starts with localized activities of maternal determinants: *bicoid (bcd)* in the anterior pole, and *nanos (nos), oskar*, and *caudal* in the posterior. Anterior zygotic proteins including *hunchback (hb)* and *Krüppel (Kr)* are activated by *bcd*, while posterior *knirps (kni)* and *giant (gt)* are activated by *caudal* [7]. These four zygotic proteins as well as *tailless, hackbein* and three others are called *gap* proteins and are regulated by one another. The relation among these proteins are not completely clarified; *hb, Kr, kni, gt* are known to repress with one another in high concentration, but *Kr* and *hb* activate *kni* and *Kr* respectively in low concentration [3, 9, 10]. These fine *gap* tunings are the result of cooperative regulation of proteins.

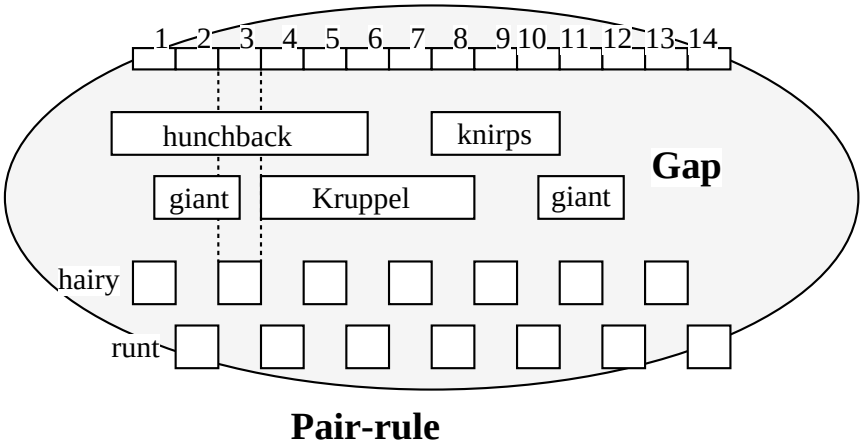


Figure 1: Segmentation

There are two steps in making *pair–rule* pattern. Primary genes including *even–skipped (eve), hairy, runt* are regulated by *gap* proteins, and are responsible for the secondary genes including *fushi–tarazu*. Here, two protein pairs, *hairy* against *runt* and *eve* against *fushi–tarazu*, complementarily emerge and form 7 stripes each. These pairs are expected to repress mutually.

The regulation mechanism is little clear [2, 6, 8], but these observations show primary *pair–rule* genes are cooperatively activated or repressed by neighboring *gap* proteins. What we want to know is following aspects.

- Are gradients of only *gap* proteins enough for the primary *pair–rule* activation?
- Does each stripe has its own enhancing region?

3 Related Work

Reinitz and Sharp [5] have done a good work on the same subject; they used differential equations and determined the parameters among *gap* and *pair–rule* proteins by simulated annealing.

Their prediction contains repressing and activating power, as well as diffusing, decaying, and producing rate of proteins. After having computed quantitative biological data, they drew qualitative relationship among proteins. For example, they found that *even–skipped (eve)* protein hardly diffuses.

Our approach is different though our aim is the same. We assume qualitative result can be drawn from qualitative input. As for the previous example, the fact *eve* protein hardly diffuses is clear from the sharpness of 7 stripes. (If it should diffuse, the borders of stripes must blur.) We uses only qualitative data: spatial relationship among proteins. With this abstraction, however, we have to give up simulating temporal order of stripe formation and the width of stripes. Though there are drawbacks that diffusing and producing rate are ignored, we believe that still much to see remain in the qualitative result.

4 Simfly2

In SIMFLY [1], relation among proteins, activation or repression, were specified in advance, and the simulator searched intensities of regulators (threshold values) only. The vast search space hindered us to check every regulatory relation. To overcome this problem, SIMFLY2 employs GA. Abstraction of proteins is almost the same as SIMFLY. Production rate is fixed to 1 unit for all the proteins. Diffusion rate is fixed for each hierarchy: *maternal–effect, gap*, and *pair–rule*. For efficiency, new version has decaying rate (no "lifetime" as in SIMFLY); in each time step for all the proteins, 20 % of the total amount is deleted. This shortcut does not change the maximum amount of protein in SIMFLY. We confirmed, by checking each pattern produced in SIMFLY2, that this shortcut made little change in the simulated pattern.

4.1 Threshold Model

Each protein is activated by, repressed by, or not related with other proteins. These relations are represented with two inequalities as in described in Introduction. Two threshold values range integers from 0 to 5. An example for *gap* proteins are shown below.

Rule 1

```
             bcd  caudal   hb    Kr    kni  giant
[[hb 25 4 [[(A 2) (N .)   (N .) (R 3) (N .) (N .) ]]]
 [Kr 25 4 [[(N .) (R 3)   (N .) (N .) (R 4) (R 2) ]]]
 [kni 25 4[[(N .) (S 4 1) (R 3) (R 4) (N .) (N .) ]]]
 [gt 25 4 [[(N .) (N .)   (N .) (R 4) (R 2) (N .) ]]]]
```

Two integers right after protein names specify diffusion constants. These values do not change in the course of simulation. According to threshold values, there are four types of relations.

`S`: Squelching. Next two integers are the threshold of both the upper and the lower limit.

`A`: Activation. The threshold of the upper limit is always 5 (max-threshold).

`R`: Repression. The threshold of the lower limit is always 0 (min-threshold).

`N`: No relation. Proteins do not interact. If the lower threshold is larger than the upper in `S` relation, corresponding two proteins do not interact, either.

A relation `(S 4 1)` in *kni* rule of *caudal* column is interpreted as

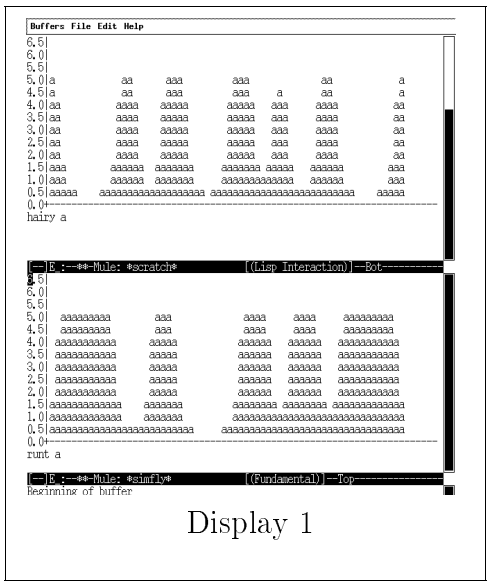
$$\textbf{if } \texttt{4} > \texttt{amount(caudal)} > \texttt{2} \textbf{ then } \texttt{create(kni)} \ .$$

From Rule 1, you can tell that *hb* emerges only when *bcd* is more than 2 and *Kr* is less than 3.

Pair–rule proteins may have multiple regulatory loci. Each relation in a rule is interpreted as conjunct, and each juxtaposed rule as disjunct.

$$\texttt{[name [[(S1) (S2) (S3)]} \\ \texttt{\ \ [(S4) (S5) (S6)]]]} \quad \leftrightarrow \quad \begin{array}{l} \texttt{create(name) if} \\ (\ \mathrm{S1} \wedge \mathrm{S2} \wedge \mathrm{S3}\ ) \vee \\ (\ \mathrm{S4} \wedge \mathrm{S5} \wedge \mathrm{S6}\ ) \end{array}$$

For example, assume protein *hairy* and *runt* have 3 regulatory loci each. Following is an example for these *pair–rule* proteins.



Display 1

Rule 2

```
[hairy 10 1 [
[(N .) (N .) (N .) (N .)   (S 5 4) (N .)   (N .) (N .)]
[(N .) (N .) (N .) (N .)   (N .)   (S 3 1) (N .) (N .)]
[(N .) (N .) (N .) (S 4 2) (N .)   (N .)   (N .) (N .)]]]
 bcd  caudal  hb    Kr      kni    giant   hairy runt
[runt 10 1 [
[(N .) (N .) (N .)   (N .) (S 4 2) (N .)   (N .) (N .)]
[(N .) (N .) (N .)   (N .) (N .)   (S 4 3) (N .) (N .)]
[(N .) (N .) (S 4 2) (N .) (N .)   (N .)   (N .) (N .)]]]
```

The expression pattern with this rule is shown in Display 1. Note that there is no threshold of 0. No threshold of 0 as the lower limit means every protein must have at least one activator; no relation can be written only with repression.

4.2 Trimming Branches

Rule 3

```
[ID EVAL [[hb 10 1 [[(A) (R) (N) (R) () () ]]]
          [Kr 10 1 [[() () () (N) (R) (R) ]]]
          [kni 10 1 [[() () (R) (R) (N) (R) ]]]
          [gt 10 1 [[() () () (R) (R) (N) ]]]]]
```

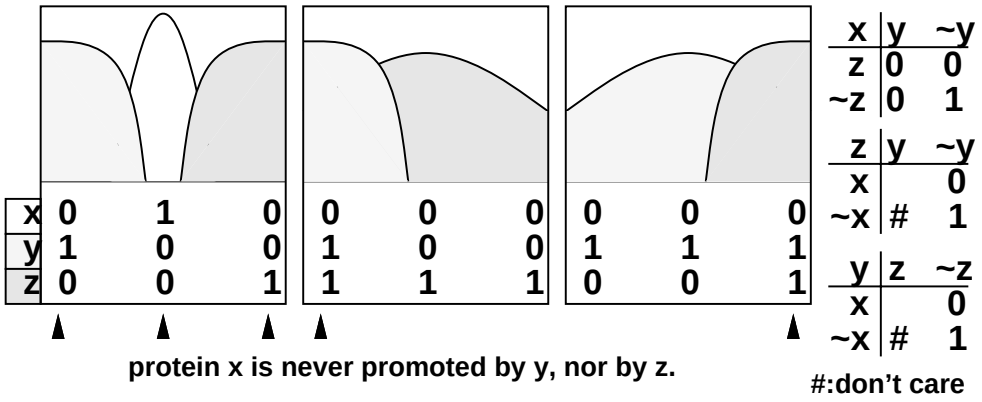


Figure 2: Guessing Relations

From observations in genetics, we can logically guess the possible relation among proteins. From expression patterns reported in literature, many relations are determined to be never repressed, never promoted, or not related.

These restrictions can be used as in Rule 3. Blank braces are regarded as `S` relation, in which two thresholds are required, while restricted braces requires only one or no threshold.

4.3 Genetic Algorithm

SIMFLY2 repeats the following cycle.

1. Create initial population.
2. Evaluate each individual and sort the population.
3. Leave better individuals and kill others.
4. Duplicate individuals, then crossover and mutate them.(Figure 3)
5. Goto step (2)

GA can search for pseudo-optimal values for a given problem, though its mechanism of performance has not yet fully analyzed. GA is a variant of hill–climbing method, and if the climbing terrain contains many sheer cliffs, it does not well work. The terrain of the problem depends on evaluation function, so what is important is definition of evaluation function. Evaluation function in SIMFLY2 is the same as that in SIMFLY; the number of matching interval relation of proteins. If the matching number of two rules are the same, the rule with less valid relations has a better score. (Remember that simpler rules are better.) With this evaluation function, the terrain still have cliffs; a single change in a rule may reduce the number of matching score by 3 or 4. Therefore, GA often fails to find the best answer. But note that GA is not an optimizing algorithm. It suffices if GA finds a new interaction among proteins in a several times of runs.

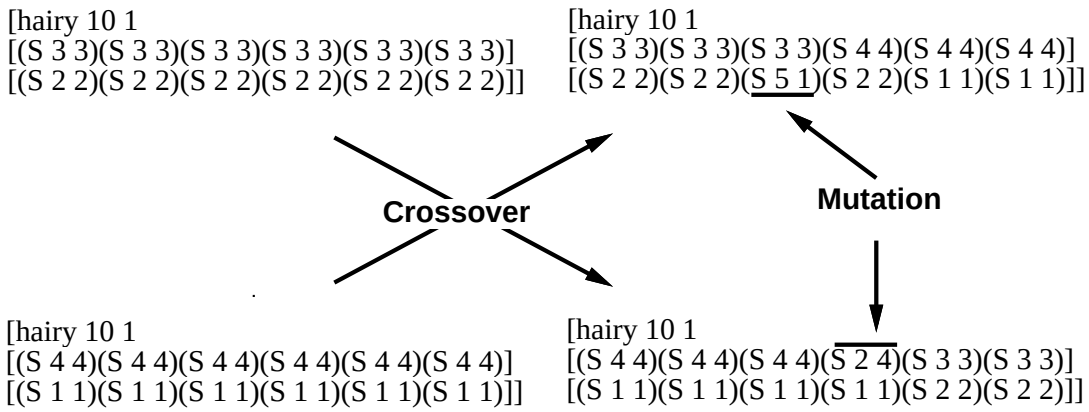


Figure 3: GA operators

4.4 Result

First we tested the rule of previous SIMFLY, and confirmed that GA did find the optimal relations we had found using exhaustive search in the last version.

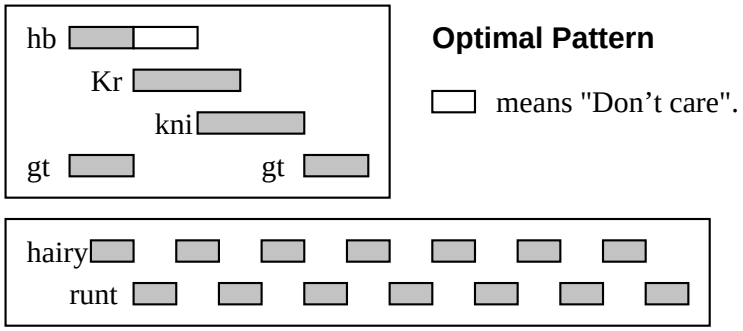


Figure 4: Optimal Pattern

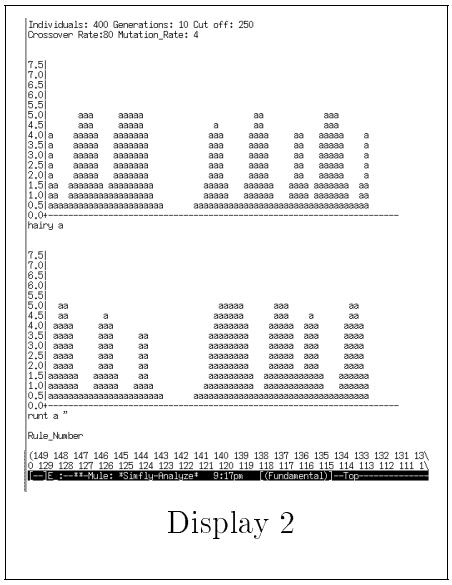
Rule 4

```
             bcd   caudal  hb    Kr   kni   gt
[[hb 10 1 [[(A 13)(R 1235)(N .) (R 3)(N .) (N .)  ]]]
 [Kr 10 1 [[(N .) (R 3)   (N .) (N .)(R 45)(R 13) ]]]
 [kni 10 1[[(N .) (S 4 1) (R 34)(R 4)(N .) (R 145)]]]
 [gt 10 1 [[(N .) (N .)   (N .) (R 4)(R 13)(N .)  ]]]]
```

Started with the total population 300, mutation rate 5%, crossover rate 80%, and survivors' population 40 for each generation, GA converges within 20 generations to Rule 4. For `R` and `A`, only one threshold is shown. (The other is fixed.) Juxtaposed numbers mean that the threshold can take any of these numbers. Note that GA does not always find this result.

From this result, we can tell that repression of *kni* by *gt* and that of *hb* by *nos* are redundant. See the threshold takes from 1 to 5. We can draw Rule 1 from this result. For each run of GA, we can find a new regulatory patterns in this way. Some other patterns are shown in Appendix.

Next, we tested how *pair–rule* stripes can emerge from the optimal *gap* protein pattern. With relations for *gap* proteins fixed, 3 rules for *hairy* protein and 3 for *runt* were searched. If we do not allow 0 to be a threshold, Rule 2 are produced. If we allow, Rule 5 is produced. The expression pattern with this rule is shown in Display 2.



Display 2

Rule 5

```
[hairy 10 1 [
[(N .) (N .) (N .) (R 4) (R 2) (R 3) (N .)(N .)]
[(N .) (N .) (N .) (N .) (N .) (A 4) (N .)(N .)]
[(N .) (N .) (N .) (R 2) (N .) (R 1) (N .)(N .)]]
[runt 10 1 [
[(N .) (N .) (N .) (N .) (N .) (N .) (R 1)(N .)]
[(N .) (N .) (N .) (A 3) (N .) (A 2) (N .)(N .)]
[(N .) (N .) (N .) (R 3) (A 4) (N .) (N .)(N .)]]
```

Note that both proteins contain many `N` rules. Simpler rules seem to emerge easier than complex ones. In Rule 5, *runt* does not regulate *hairy*. More closer look reveals that as much as 5 *hairy* stripes emerge through only repression by *gap* proteins (see the first rule for *hairy*). With repression only, many stripes can hierarchically emerge in this way, but it is unlikely that a protein has no activator. With activation, however, neither hierarchical regulation nor mutual interaction between *pair-rule* proteins is observed as in Rule 2.

5 Discussion and Future Work

5.1 GA Parameters

In GA, most important is the parameters set for each problem: crossover rate, mutation rate, and initial population. In fact, without knowledge of terrain of a given problem, we cannot properly determine these values. Currently the population is set to 300 $\sim$ 400. Larger population is better, but we need to make the program more efficient to simulate with more population. Implementation on a more efficient machine is our future work.

5.2 Pair–rule Proteins

Primary pair–rule proteins, *(eve, hairy, runt)*, interact with one another. We have simulated only *hairy* and *runt* protein, but other proteins need to be simulated to find the real interaction.

With 2 proteins only, however, the number of threshold values exceeds 40. To cope with more proteins, the simulator needs to be much more efficient.

6 Conclusion

We propose threshold model for qualitative simulation of proteins. After this model, SIMFLY2 is implemented for the simulation of DNA-binding proteins in *Drosophila* embryogenesis. Using GA, SIMFLY2 found the relation among *gap* proteins, interaction which SIMFLY found in our last work. SIMFLY2 also suggests following relation among *gap* and *pair–rule* proteins.

- *Pair–rule* proteins are regulated by a small number of proteins.
- *Pair–rule* 7 stripes emerge in hierarchical manner, if a protein may not have its own activator. If the protein must have its activator, stripe region is simultaneously specified through subtle changes in threshold values.

Acknowledgment

We thank Prof. Iba for his helpful comment and his suggestion about GA. We also thank Prof. Takagi at Human Genome Center (HGC) for letting us use the facilities in HGC. This work was supported by a Grand-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] M. Arita, M. Hagiya, and T. Shiratori. Geisha system: an environment for simulating protein interaction. In *Proceedings Genome Informatics Workshop 1994*, pages 80–89, 1994.

[2] J. A. Langeland and S. B. Carrol. Conservation of regulatory elements controlling hairy pair-rule stripe formation. *Development*, 117:585–596, 1993.

[3] M. J. Pankratz, M. Hoch, E. Seifert, and H. Jackle. Kruppel requirement for knirps enhancement reflects overlapping gap gene activities in the drosophila embryo. *Nature*, 341:337–339, 1989.

[4] M. Ptashne. *A Genetic Switch*. Blackwell Scientific Pub., 1987.

[5] J. Reinitz and D. Sharp. Mechanism of eve stripe formation. Technical Report LA-UR-94-1915, Los Alamos National Laboratory, 1994.

[6] G. Riddihough and D. Ish-Horowicz. Individual stripe regulatory elements in the drosophila hairy promoter respond to maternal, gap, and pair-rule genes. *Genes & Development*, 5:840–854, 1991.

[7] R. Rivera-Pomar, X. Lu, N. Perrimon, H. Taubert, and H. Jackle. Activation of posterior gap gene expression in the drosophila blastoderm. *Nature*, 376:253–256, 1995.

[8] S. Small, R. Kraut, T. Hoey, R. Warrior, and M. Levine. Transcriptional regulation of a pair-rule stripe in drosophila. *Genes & Development*, 5:827–839, 1991.

[9] D. Stanojevic, T. Hoey, and M. Levine. Sequence-specific dna-binding activities of the gap proteins encoded by hunchback and kruppel in drosophila. *Nature*, 341:331–335, 1989.

[10] J. Treisman and C. Desplan. The products of the drosophila gap genes hunchback and kruppel bind to the hunchback promoters. *Nature*, 341:335–337, 1989.

Appendix

Another rule for *gap* proteins

Protein *hb* is known to be repressed by *nos*, but in our simulation, activation by *bcd* is sufficient for its localized expression.

```
 [hb 10 1 [[(A 3) (N .) (N .) (N .) (N .) (N .)]]]
 [Kr 10 1 [[(R 3) (R 3) (N .) (N .) (R 3) (R 4)]]]
[kni 10 1 [[(N .) (A 1) (N .) (R 4) (N .) (R 1)]]]
 [gt 10 1 [[(N .) (N .) (R 4) (R 1) (R 4) (N .)]]]
```

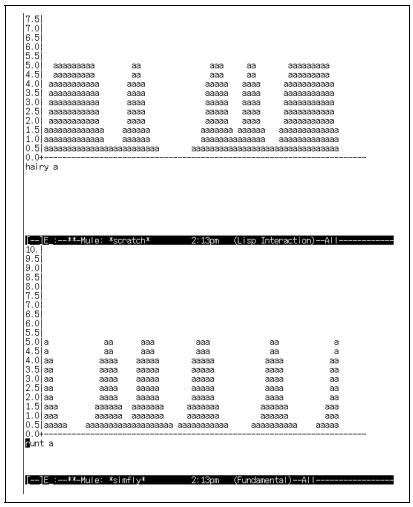
Another rule for *pair–rule* proteins
(Activation not always required. Display 2)

Hierarchical regulation is clear in this rule. First *hairy* emerges in and from between *gap* proteins. Next, *runt* comes out where there is no *hairy*.

```
[hairy 10 1
[[(N .) (N .) (N .) (R 3) (R 3) (R 3) (N .) (N .)]
 [(N .) (N .) (N .) (N .) (N .) (A 4) (N .) (N .)]
 [(N .) (N .) (N .) (N .) (A 4) (N .) (N .) (N .)]]
[runt 10 1
[[(N .) (N .) (N .) (N .) (N .) (N .) (R 1) (N .)]
 [(N .) (N .) (N .) (N .) (A 3) (N .) (N .) (A 5)]
 [(N .) (N .) (N .) (N .) (N .) (N .) (N .) (A 3)]]
```

Another rule for *pair–rule* proteins
(Activation always required.)

This produces similar result as that of Rule 2. There is no hierarchical interaction between *pair–rules*.



```
[hairy 10 1 [
[(N .) (N .) (N .) (N .)   (N .)   (S 4 3) (N .) (N .)]
[(N .) (N .) (N .) (S 3 1) (N .)   (N .)   (N .) (N .)]
[(N .) (N .) (N .) (N .)   (S 4 3) (N .)   (N .) (N .)]]]
[runt 10 1 [
[(N .) (N .) (N .) (N .)   (N .) (N .)   (N .) (N .)]
[(N .) (N .) (N .) (N .)   (N .) (S 3 1) (N .) (N .)]
[(N .) (N .) (N .) (S 4 2) (N .) (N .)   (N .) (N .)]]]
```