Chapter 1
Transcription Dynamics: Cellular Automaton Model of Polymerase Dynamics for Eukaryotes

Yoichi Nakata, Yoshihiro Ohta, and Youichiro Wada

1.1 Brief Review of Transcription

The cells of living organisms make proteins to reproduce themselves and express various functions in their activities. The making process of proteins is divided into two stages. First, a molecular motor called RNA polymerase (in eukaryotes, three types of RNA polymerases and RNAPII are in charge) recognizes an amino acid code described by deoxyribonucleic acid (DNA) and duplicates it using ribonucleic acid (RNA), resulting in the generation of messenger RNA (mRNA). This process is called transcription. Next, newly synthesized mRNA is decoded by ribosome, and a sequence of amino acid, namely protein, will be generated (Fig. 1.1).

To respond to both outer and inner stimuli, transcription of required protein coding gene will get started, terminates once the necessary amount of the proteins has been created, and the target protein depends on the cell type and situation. In the previous notion, genomic DNA was considered to move very flexibly during transcription process. This is because DNA is a linear polymer of alternating phosphates and sugars, each of which is bound to one of the four types of bases—20 adenine, cytosine, guanine, and thymine (Figs. 1.2 and 1.3). Triad of nucleic acid corresponds to a single amino acid, and this is called as codon. In humans, DNA consists of about 6 billion deoxyribose (i.e., it has 2^{12*10^9} bits of data). But only a tiny percentage of them are used [5].

Y. Nakata (✉) · Y. Wada
Isotope Science Center, The University of Tokyo, Tokyo, Japan
e-mail: ynakata@ric.u-tokyo.ac.jp; wada-y@lsbm.org

Y. Ohta
Arithmer Inc., Tokyo, Japan
e-mail: ohta@arithmer.co.jp

T. Tokihiro (ed.), *Mathematical Modeling for Genes to Collective Cell Dynamics*,
Theoretical Biology, https://doi.org/10.1007/978-981-16-7132-6_1

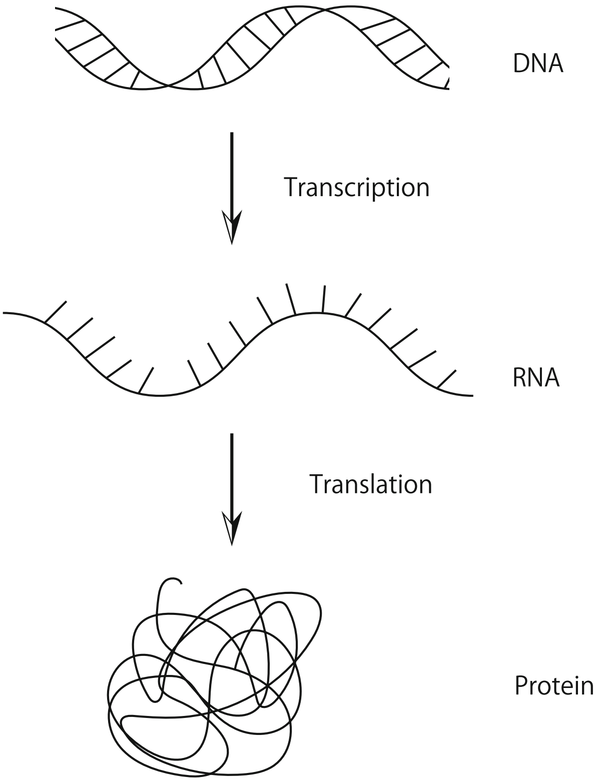


Fig. 1.1 Picture of central dogma. This picture is from https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet

If we focus only on the base information, we do not determine which terminal should be selected as the start point, so it is necessary to determine the order. In the sugar, each carbon atom is labeled as in Fig. 1.2, and the 5' and 3' carbons are involved in polymerization. The terminal deoxyriboses have 5' or 3' carbon unused for polymerization, which is called 5'-terminal and 3'-terminal, respectively. The direction of the sequence is determined by the order from 5'- to 3'-terminal. This is also the orientation of the electrical polarity of DNA. However, it is noted that this orientation does not always coincide with the actual direction of RNAP reading DNA during transcription. In a single spice, genomic DNA has an identical sequence among all the cells and individuals (only immune response-related cells are exceptions). For chemical stability, the DNA forms a double helix structure by hydrogen bonding with the complementary strand determined by the complementary relationship of A-T and G-C. Therefore, DNA sites are counted by

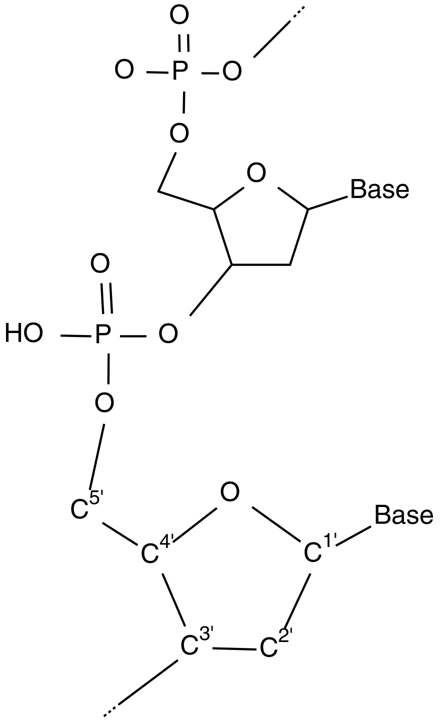
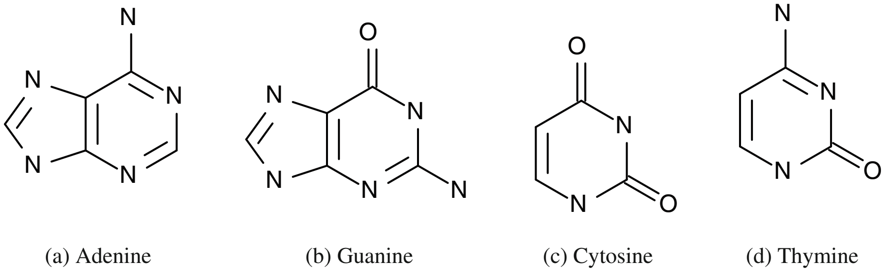


Fig. 1.2 Chemical structural formula of DNA



(a) Adenine (b) Guanine (c) Cytosine (d) Thymine

Fig. 1.3 Structures of four nucleotides

base pairs. The DNA double helix is stored in a form that wraps around protein octamers called histone. A set of this histone and the wrapping DNA is called a nucleosome, and a structure of compacted nucleosomes is called chromatin. Among chromatins, that of tightly bonded by nucleosomes are called heterochromatin regions, and conversely, that of loose bonds are called euchromatin regions. In transcription, euchromatin regions are prone to accept active RNAPIIs.

During cell division, chromatin gathers to form a pair of rods called chromosomes. The number of chromosomes varies depending on the organism, but in

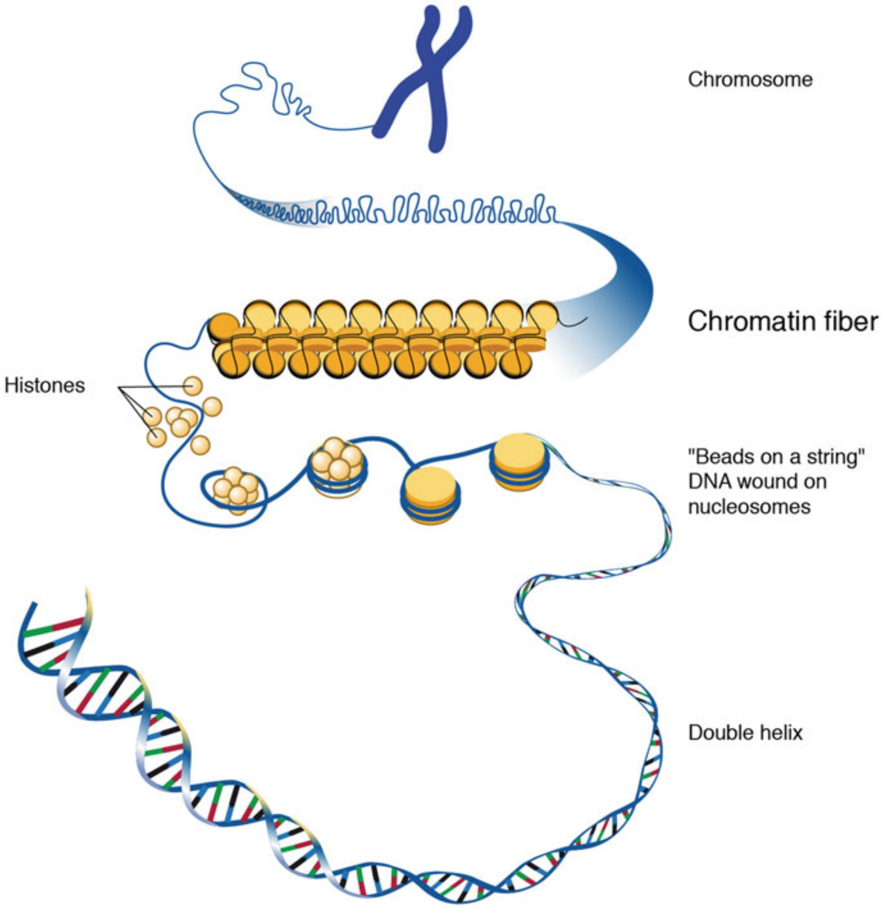


Fig. 1.4 A hierarchy of DNA. (This picture is a modified version of the one from https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics)

humans, it is known that there are typically 23 pairs of chromosomes. The position of a base pair in the sequence of DNA is expressed by using chromosomes (Fig. 1.4).

Only a few parts of DNA carry protein sequence code, and it is only 1.5% in humans. DNA regions involved in the production of specific proteins are called genes. Approximately, 26,000 genes are known in humans. Furthermore, in eukaryotes, there are also DNA regions within genes that are independent of the sequences of the proteins produced along the way. This region is called intron, and conversely, the region coding protein sequence is called exon. The location of genes, exon, and intron and the location of chromosomes are specified and databased.

RNAP generates RNA based on all the information in the gene region regardless of introns or exons, and this preliminary product is called pre-messenger RNA.[1] From this product, the parts derived from the intron regions are spliced out at some timing (in many cases, it is thought that RNAP reaches the exon region from the intron region). Eventually, the RNA part originated from the exon regions will be composed, and this is called messenger RNA (mRNA). This editing procedure is called splicing. Next, the mRNA is chemically treated at both ends so that it can exist outside the nucleus in order to generate proteins. This is where a straightforward question emerges. Why is there an inefficient part like intron? It seems to be good that genes hold only the minimum information necessary to make proteins, and RNAPII reads the only such codes to create mRNA. One hypothesis is that, when splicing, the intron-corresponding part and the exon part sandwiched between them are decomposed together to increase the final product variation. This phenomenon is called alternatively splicing, and the mRNA in which a part of the generated exon disappears is called a splice variant. Besides, there are research reports that the length of introns affects the rhythm of transcription, and as a result, the morphology of organisms is determined by regulating expression [16].

The information described using four types of sequences A, T, G, and C is generically called a genome. By 2003, the Human Genome Project revealed all human genome information. Therefore, anyone can access genome information from the database. At first, it was thought that all life phenomena could be known entirely if the genome was known, but it gradually recognized that there were phenomena that could not be explained by genome information alone.

For example, a queen bee breeds both a worker one and the next generation queen one, but there is no genomic difference between them. It is known that only those who have been given a special meal (royal jelly) in childhood can become queen bees. In addition, for humans born to mothers who became hungry during pregnancy, there was a statistically significant difference in the prevalence of lifestyle diseases when children grew up, depending on when hunger occurred during pregnancy. What is happening to their bodies? Several experiments reveal that transcription (and function expression) is affected by the replacement of a specific hydrogen group with a methyl group or an acetyl one in the base moiety part of the DNA site or histone tails (protein chains growing from each histone). For example, transcription is suppressed by cytosine methylation of DNA. When a specific one of the hydrogen group in the 27th amino acid residue in histone tail of H3 histone (which is known to be lysine) is replaced with a trimethyl group, DNA is firmly wrapped around a histone, making DNA hard to be involved in transcription. Conversely, methylation on the 27th amino acid of H3 histone is replaced with an acetyl group, DNA wrap is weakened, and transcription enhanced. Note that these modifications happen not in the genome itself because these modifications do not change DNA sequences. However, DNA modification pattern

[1] mRNA is that finally produces proteins. Other RNAs exist and are thought to play an important role in controlling transcriptional dynamics. However, that is not covered in this book.

is thought to be as crucial as genomic information because it significantly impacts transcription. Such a factor affecting transcription other than the genome sequence itself is called an epigenetic information, namely epigenome. All of the previous examples are regarded as epigenomic effects. It is known that the epigenome retains its information during cell division, and the fertilized egg inherits the parent epigenome, but most of the egg is reset just before the egg begins to differentiate into various cells. However, it has been found that some of the epigenome information is not reset by differentiation and is inherited to offspring as a result [23].

From here, let us consider the rough behavior of RNAP during transcription (RNAPII in the case of eukaryotes). This is a complex composed of 12 subunits in the nucleus and binds to a promoter region called TATA box motif (literally including a repeated TATA sequence), locating about 5 kbp upstream of the gene, and RNAP drifting in the nucleus binds to TATA box binding protein. Both proteins attract each other, and RNAP finally attaches to the DNA site and begins to move toward the transcription starting site. After arriving, it starts transcription and moves on the DNA track while generating pre-mRNA by reading the base information of DNA. This transcription direction depends on the gene and does not necessarily match the DNA direction. Sometimes RNAP may move in the opposite direction with DNA. RNAP terminates transcription when it arrives at the end of the gene region and desorbs. RNAP again drifts in the nucleus after it leaves the DNA and until it binds to the protein again. In this model, the movement of single RNAP is affected by the position of other RNAPs that run in front in addition to the genome sequence. This is very similar to the movement of a car on the road as described below. RNAP overtakes, slows down, and sometimes collides. This model works reasonably well for organisms with simple structures such as prokaryotes. For example, it has been confirmed that RNAP collides or slows down in prokaryotes [34, 35]. However, such a model can be applied to lives with simple structures like prokaryotes. It is known that the dynamics of RNAPII in eukaryotes, including humans, behave more complicated due to the reasons as follows. First, the genes of eukaryotes include not only protein-encoding regions (exon) but also introns. The pre-mRNA regions synthesized from intronic regions degrade rapidly and do not leave in the mRNA, which is the final product of transcription. Second, several proteins control the dynamics of RNAPII. Such proteins combine specific sites on DNA tracks and prevent RNAPII movement physically or help RNAPII attach to the transcription start site or attach RNAPII for starting transcription [14, 19]. There is a relation between a specific epigenome modification for some DNA sites and attachment of corresponding transcription factors [12, 17, 20, 21]. Finally, it is speculated that RNAPII does not exist alone, but a huge protein complex includes several RNAPII. The interactions that mediate this complex can also occur over long distances in the genome coordinates [5, 8–10, 29].

In order to create a model that describes the dynamics of RNAPII, we must take these into account. In the following sections, we will explain what experimental results have been reported on transcriptional dynamics and what dynamics we make up based on that.

1.2 Experimental Result

In this section, we will briefly explain what RNAPII behavior during transcription has been known by several experimental results before modeling [19, 33, 40]. To develop an RNAPII transcriptional dynamics model, we need to know the location of RNAPII. As one method, Wada and Ohta et al. investigated where and how much RNAPII was currently at some time by measuring the copy number of RNA fragments, which were recollected from living cells, labeled by fluorescent dyes, and quantified by 25mer nucleotide probes. This method is called a tiling array.

Formerly, we delivered Tumor Necrosis Factor-alpha (TNF-α), one of the inflammatory cytokines, to human umbilical vein endothelial cells (HUVECs) and observed the movement of the generation of nascent RNA by inflammation stimulation. Based on the previous experiment, we focused on long five genes, like the SAMD4A gene region for observation because it has good inflammation responses and sufficient length for observing RNAPII dynamics.

Visualization of nascent RNA generation was done in the specific genes at intervals of 7.5 min from 0 to 180 min after stimulation (Fig. 1.5). By analyzing the data, the following was obtained:

Transcription is not constantly performed after stimulation, and there are periods during which transcription is actively performed, such as waves. It takes about 15 min from stimulation to the first transcription.

At the transcription start point, constitutive active transcription happens, but many will be terminated before reaching to the end of genes. There is something like a checkpoint that shuts down RNAP transcription until a stimulus is activated to activate it.

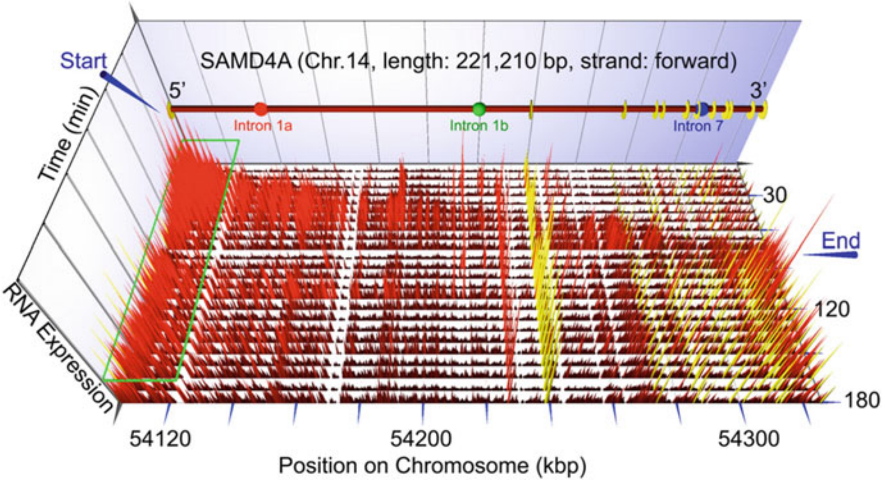


Fig. 1.5 Time evolution of the distribution of RNAPII on SAMD4A gene

When there is RNAPII performing transcription in the front, other RNAPIIs near the rear do not perform transcription.

By examining five representative genes, the transcriptional wave appears to be traveling at an average speed of approximately 3.1 kb per minute. It does not move at a uniform speed everywhere but depending on the region. For example, it seems to go faster in introns and slower in exons. This result has also been reported by Kolasinska-Zwierz et al. [19] and Schwartz et al. [33].

At the DNA site where a protein called CTCF is bound, the movement speed of RNAPII becomes slow. This is because CTCF might physically hinder RNAPII, this is why CTCF localizes the RNAPII migration range. When CTCF was knocked down, the movement of RNAPII was spread, and it is observed to happen at a place where transcription should not be expected. It is already known that the presence of CTCF is confirmed at the boundary of the active transcription region.

Looking at the RNA density profile, the density does not increase or decrease continuously from the beginning, but an isolated peak at a specific DNA site (around 0) appears over time. Most of those places are in exons, so it seems that it can be explained by the intron–exon speed ratio, but for that, you have to set a very high-speed ratio. The fact that the first intron does not have such a speed ratio cannot be ignored.

Recently, it should be noted that next generation sequencer allows us to detect the position efficiently by identifying the sequences of nucleotide fragments bound and recollected from specific proteins, including RNAP. The detected localization of RNAPII was consistent to that of transcription wave.

Further, the authors have reported that the observed "wave of transcription" was well explained by the presence of a complex called a transcription complex composed of RNAPII, which had been predicted to exist. The transcription wave was considered to be generated by the rapid change of the chromatin structure, which might be caused by transcription complex, the so-called transcription factory. Here, a mathematical model is constructed based on some of these facts in the following sections. First, we will explain a mathematical concept called cellular automaton as a tool.

1.3 Cellular Automaton and Traffic Flow Model

1.3.1 What Is Cellular Automaton?

A cellular automaton (CA) is one of the discrete dynamical systems. It has discretized states in lattices called cells and updates states of the cells for each discretized time. In other words, it is the dynamical system in which all dependent and independent variables (space, time, and states) are discrete. Ulam and von Neumann first propose the cellular automata concept to solve the self-reproduction problem of lives. The cellular automata represent complex nonlinear phenomena even if their update rules are simple.

The elementary cellular automaton (ECA) is a class of cellular automata with one-dimensionally configured cells with only two states (0 or 1), and the update of a cell is determined by itself, one before and after. That is, by denoting $t \in \mathbb{Z}_{\geq 0}$, $j \in I$ ($I \subset \mathbb{Z}$ is a discrete interval), and $u_j^t \in \{0, 1\}$ as the time, the site number and the state of site number j at time t, respectively, under the assumption of the initial condition $\{u_n^0\}_{n \in I}$ and the boundary condition, the time evolution of $\{u_j^t\}$ is expressed as $u_j^{t+1} = f(u_{j-1}^t, u_j^t, u_j^{t+1})$ with a function $f : \{0, 1\}^3 \to \{0, 1\}$.

Since an ECA is regarded as a partial-difference equation, one needs to determine a proper boundary condition. Most of the cases, one assumes the Dirichlet boundary condition $u_{j'}^t = 0$ (j' is a boundary of I) or the periodic one $\exists N, u_{j+N}^t = u_j^t \; \forall t \geq 0$.

The amount of functions f (the update rules of ECA) is $2^{2^3} = 256$. We identify them by integers $k := \sum_{i,j,k=0,1} f(i, j, k)2^{2^2 i+2j+k}$. The ECA with integer index M is called ECA Rule M or ECA M. The behavior of ECA was actively analyzed, for example, by Wolfram [43]. There are several classifications for them, but the most common one is as follows [42]:

Class I: Uniform
Class II: Periodical
Class III: Chaotic
Class IV: Complex

The ECAs in classes III and IV show interesting behavior and sometimes exhibit characteristics of a real system.

An interesting pattern may be drawn when states of cellular automata are properly arranged at each time and space. For example, the time evolution of ECA Rule 90 with an initial state where only one point is 1 and the others are 0 is shown in Fig. 1.6. The time-space pattern is similar to a fractal pattern called a Sierpinski gasket.

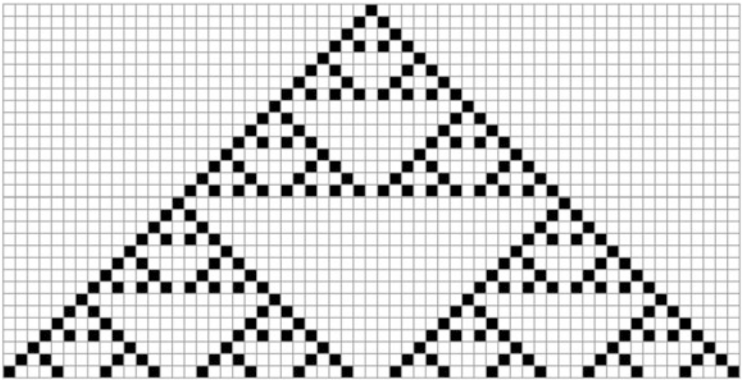


Fig. 1.6 An example of time-space pattern of ECA Rule 90. Black boxes mean $u_j^t = 1$ for these sites and white ones mean $u_j^t = 0$

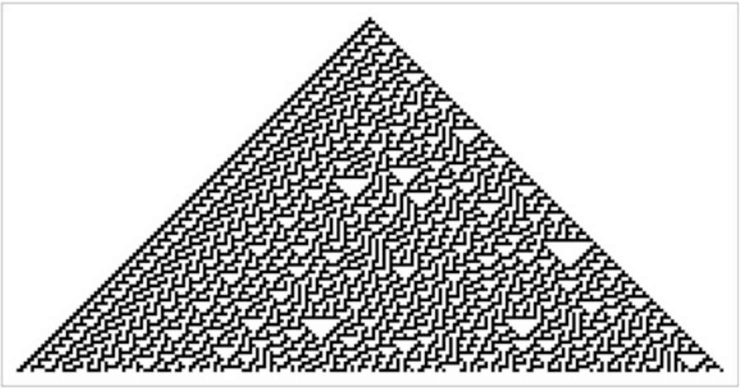


Fig. 1.7 An example of time-space pattern of ECA Rule 30. Triangle patterns seem to appear randomly

Figure 1.7 shows the space-time pattern of ECA Rule 30 with the same initial condition. It draws a similar fractal pattern but a more complex one. Interestingly, this pattern is very similar to that of a shell of the conidae. The question of why creatures create designs identical to ones in a simplified mathematical model is a fascinating subject, but it is beyond the scope of this book, so we will stop explaining here.

We can extend the elementary cellular automata by setting the cell configuration from a one-dimensional lattice to a higher-dimensional one or expanding the range of cells to be referenced on an update. We can express more complex natural phenomena for such extended models, such as solitons (the solitary waves that preserve their shapes after collision). One of the most famous examples is the Conway's life game, whose cells are configured on the two-dimensional lattice (with a proper boundary condition) and take two states—alive or dead. The state of a cell is updated using the following rules depending on the state of itself and those of its eight neighboring cells (Fig. 1.8):

- If the cell is alive, the next state is alive only if there are just two or three cells alive.
- If the cell is dead, the next state is alive only if there are just two cells alive.
- Otherwise, the next state of the cell is dead.

This rule indicates that life cannot survive alone but requires the cooperation of others. However, it also cannot survive if there are too many others around because they compete for limited resources. This exquisite balance of survival and death conditions creates a variety of patterns (Fig. 1.9).

Conway's life game (as well as most general cellular automata as a whole) is a good target for programming practice. It is a good study to write a program emulating the time evolution of the system by yourself. We will cite Golly (http://golly.sourceforge.net) as a ready-made program for the life game (Golly itself can

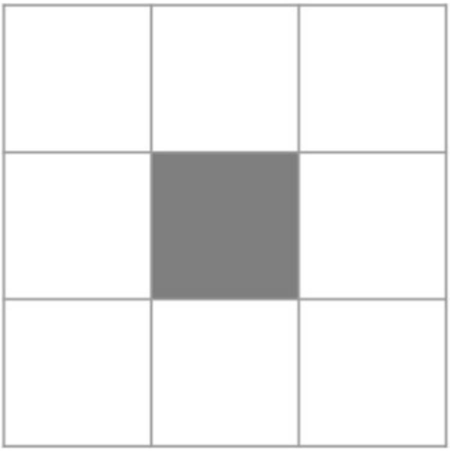


Fig. 1.8 The next state at the gray cell in the center is determined by eight surrounding neighbor sites in the figure

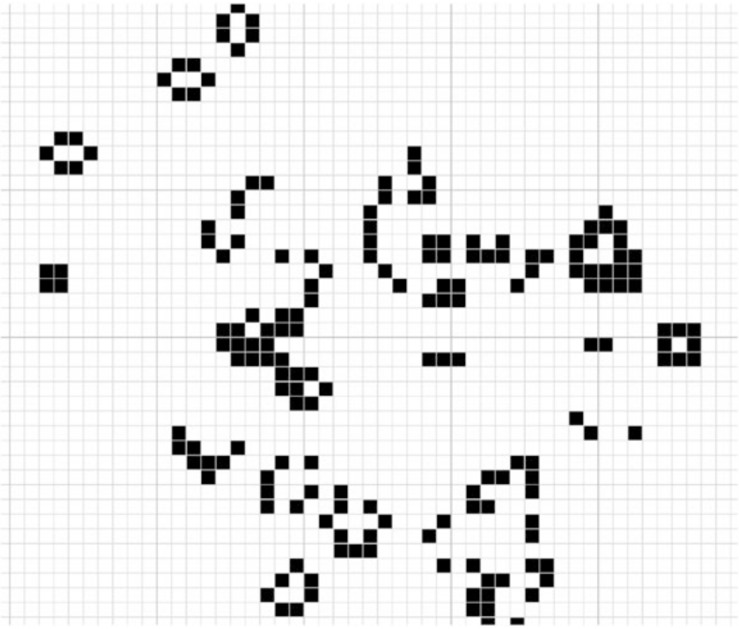


Fig. 1.9 A state of Conway's life game. The black cells mean alive and white cells mean dead

calculate time evolutions of more cellular automata). Anyway, when you have a program that describes the development rules for life games in some form, you can confirm that the system shows really fertile behavior when you start the simulation with appropriate initial values.

Due to this time evolution rule, the life game makes patterns that change periodically, translate while changing their shapes periodically, or regularly generate such parallel movement ones. By setting the initial state properly, one can express logic gates. Therefore, the life game can be a universal Turing machine.

Unlike the models using differential equations, it is very suitable with computer simulation because there is no discretization error or numerical error in simulation. It is possible to calculate quickly for improving calculation accuracy is unnecessary. Then, cellular automata models are often adopted to reproduce natural phenomena by computer simulations. One of the typical examples is the traffic flow model described in the next section.

Finally, the relationship between the phenomena by the CA and the similar ones by the differential equations is an interesting problem. Actually, this is one of the questions suggested by Wolfram. As one of the answers, it is known that we can derive the time evolution equation of a CA from a well-discretized differential equation by the limiting procedure called ultradiscretization [37].

1.3.2 Traffic Flow Cellular Automaton

Now, let us go back to ECA. We focus on ECA Rule 184. We note that there is no dependency of u^t_{j-1} to determine u^{t+1}_j when $u^t_j = 1$ and $u^{t+1}_j = 0$ when $u^t_{j-1} = u^t_j = u^t_{j+1} = 0$. By these facts, we can determine the update of the system even if it has an infinite amount of sites by setting an initial condition as $u^t_j = 0$ for $|j| \gg 1$.

Then, by interpreting that the sites are sequential boxes and each box has a ball if $u^t_j = 1$ for time t and j-th box and no balls if $u^t_j = 0$, we can rewrite the time evolution rule of ECA 184 as follows (Fig. 1.10):

- All balls are moving to the orientation that increases j.
- If there are no other balls at the next box, the ball moves there.
- If there is another ball at the next box, the ball stays.
- The next state is that by applying these rules above once for all sites.

Generally, the order to apply this update rule is very important, as such models. We apply this at the same time for all balls (remember the definition of the ECA update rule). That is, a ball cannot enter the next site simultaneously even if it is going to be empty by another ball leaving. This model is known as the simplest one to express traffic jams. The rule that one box can contain at most one ball corresponds to the exclusive volume effect. The traffic jam is finally solved when the number of balls (cars) is less than half of that of sites and never solved if more.

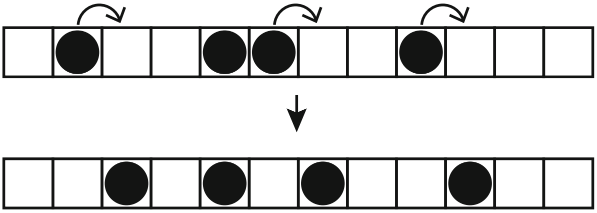


Fig. 1.10 Time evolution of ECA 184

There are several ways to express f explicitly, but the physically meaningful expression is, for example,

$$u_n^{t+1} = u_n^t + \min(1 - u_n^t, u_{n-1}^t) - \min(1 - u_{n+1}^t, u_n^t). \tag{1.1}$$

Transforming this equation, we have

$$u_n^{t+1} - u_n^t = \min(1 - u_n^t, u_{n-1}^t) - \min(1 - u_{n+1}^t, u_n^t). \tag{1.2}$$

Note that $\max(1 - u_n^t, u_{n-1}^t)$ expresses the number of moving particles from $n - 1$ to n. The right-hand side means the total change of the number of balls at j. On the other hand, the left-hand side means the time change of the number of balls. Therefore, this expression is a CA analog of the equation of continuity in fluid dynamics. ECA 184 is obtained by the limiting procedure called "ultradiscretization" from a proper discretization of the Burgers' equation, which is a differential equation model expressing traffic jam (and originally proposed to represent shock waves in the compressed fluid) [25].

Now, we introduce a stochastic factor. Let p be a given parameter satisfying $(0 < p < 1)$. Each particle moves to the next site at probability p if possible. This model is called the Totally Asymmetric Simple Exclusion Process (TASEP). Recently, the mathematical model based on TASEP is often employed to solve traffic jam problems. By introducing stochastic i.i.d. variables $\{U_n^t\}_{t \geq 0, n \in \mathbb{Z}}$ that take 1 at probability p and 0 at $1 - p$ and modifying the update rule (1.1), the time update rule of TASEP is expressed as

$$u_n^{t+1} = u_n^t + \max(1 - u_n^t, u_{n-1}^t, U_n^t) - \max(1 - u_{n+1}^t, u_n^t, U_{n+1}^t). \tag{1.3}$$

There are several ways to extend TASEP. For example, a ball can go to the second next site, or the moving probability depends on the distance of the next ball. By these extensions, we can have nearer expressions of the dynamics of real cars. For the extended CA models, a smooth traffic state becomes a metastable one when the number of cars is larger, and with a little perturbation, this state suddenly breaks, and the traffic jam finally occurs. This is considered as the mechanism of traffic jams occurring.

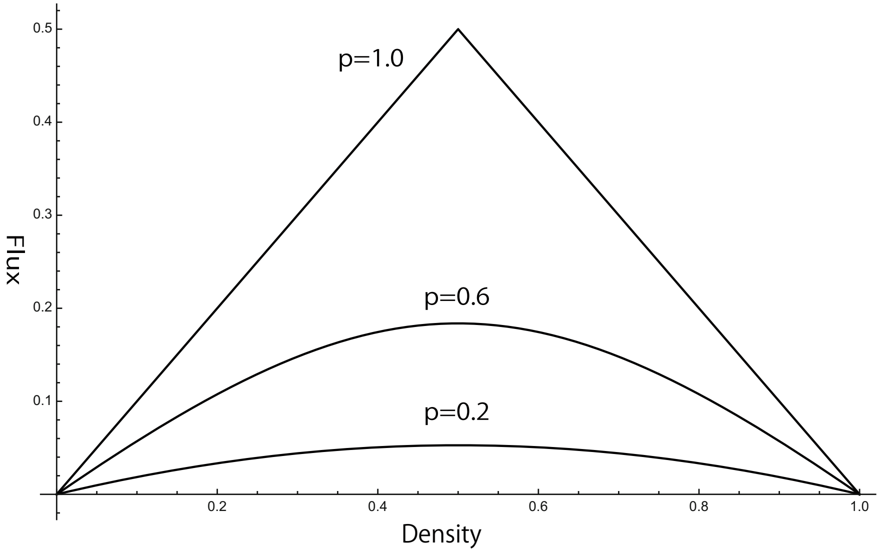


Fig. 1.11 Fundamental diagram of TASEP

Due to the observation above, the state of TASEP finally goes to a relaxed state. We define the flux of finite site TASEP as the average of moving balls per site (that is, $\sum \max(u^t_j, 1-u^t_{j+1})/N$) in the relaxed state. The graph of density of balls versus flux is called the "fundamental diagram," which is one of the most important pieces of information to understand the behavior of a model. For example, TASEP under the periodical boundary condition is expressed as Fig. 1.11. As referred before, if the number of balls is small, the traffic jam is finally solved, and all balls must move for each time evolution when sufficiently large time passed. Then, the flux is increasing monotonically depending on the density of particles. However, if the number of particles becomes larger, there are jams never solved. The number of jam trapped balls becomes larger as the density increased. Then, the flux is decreasing monotonically. In general, by denoting the number of sites and balls N and M, the density $\rho = M/N$ and the moving probability p and taking limit $M, N \to \infty$ to preserve M/N is constant, the flux is expressed as $\frac{1-\sqrt{1-4p\rho(1-\rho)}}{2}$ [30] and especially $\max(\rho, 1-\rho)$ as the limit $p \to 1$[25].

We can also set another boundary condition in which $j = 0$ and L are boundaries, and one injects a ball with probability α $(0 \leq \alpha \leq 1)$ at $j = 0$ if there are no balls, and one removes a ball with probability β $(0 \leq \beta \leq 1)$ at $j = L$ if there is a ball. Under this boundary condition, the system is known to have three phases of the flow—low density (LD), high density (HD), and flux maximized flow (MC).

There are two ways to explain the dynamics of TASEP: the view of the balls and the view of fields. Following the theory of fluid mechanics, we call the dynamics expression of the former viewpoint "Lagrangian representation" and the latter "Eulerian representation." In the infinite ECA Rule 184, denoting the position of k-th ball at time t x_k^t, one can obviously write down the time evolution rule $x_k^{t+1} = x_k^t + \min(1, x_{k+1}^t - x_k^t - 1)$. Here, it should be noted that the direction of movement is unique, and no particles overtake others. Due to these properties and the identity of the max calculation, we can directly relate these two dynamics representations [18].

1.4 Cellular Automaton Model of Transcription Dynamics

In this section, we introduce an RNAPII dynamics model by virtue of the experimental results described in section 1.2. Remember that RNAPII attaches DNA at the transcription starting site and starts transcription. Then, we coarse-grain target DNA track to a finite amount of sequential sites employ TASEP as a basic model for RNAPII dynamics on coarse-grained sites[31, 32].

We also set the stay time of RNAPII at each site [27]. The ball that arrived at a site has to stay during this time (of course, it still keeps staying after the time if there is another ball at the next site). This stay time is introduced to consider the difference of the velocity between introns and exons or due to the proteins that block RNAPII movement, for example, CTCF/cohesin (Fig. 1.12) [36].

At the given time, the ball is injected at the start site if it is empty. This means that the RNAPII free in the nucleus attaches to the start site and starts transcription.

We set a region including SAMD4A and its neighbor as the target region and the size of cell 35 bp. This size corresponds to the size of one RNAPII.

Before the numerical simulations, let us first imagine what phenomena can occur. Hereafter, for ease, we employ only the velocity difference between introns and exons and consider the semi-infinite system with the left side boundary. For such a system, we can explain the dynamics by introducing time τ_j^k as the time when k-th injected balls arrive at site j. (Note that let x_k^t be the position of k-th ball at time t, $x_k^t = j$, while $\tau_j^k \leq t \leq \tau_{j+1}^k - 1$.) Then, the dynamics is written in $\tau_{j+1}^k = \max(\tau_j^k + \gamma_j, \tau_{j+1}^{k-2} + \gamma_{j+1} + 2)$, where γ_j is the minimum stay time at site j. By using τ_j^k, the time of k-th ball spent for staying site j is $\tau_{j+1}^k - \tau_j^k$ and the time interval between when $k-1$-th ball arrives at site j and k-th ball arrives there is $\tau_j^k - \tau_j^{k-1}$. For the boundary condition, $\{\tau_0^k\}_{k\in\mathbb{N}}$ is given. For the ECA Rule 184, γ_j is identically equal to 1, which generates the trivial dynamics with no collision. To generate other states except for the free flow, one should employ the mechanism to stop the balls. In this system, the jam can occur when the latter ball at the intron site catches up with the former one at the first exon site due to the velocity gap.

Let us consider the dynamics only of $k-1$-th and k-th balls and their interval under the assumption that $k-1$-th ball never collides $k-2$-th ball since their interval is sufficiently long.

If these two balls are both in an intron and they are adjacent, the k-th one has to wait for one step until the next ball leaves. However, such a situation cannot happen unless all forward sites are occupied due to an existing jam. Therefore, under the

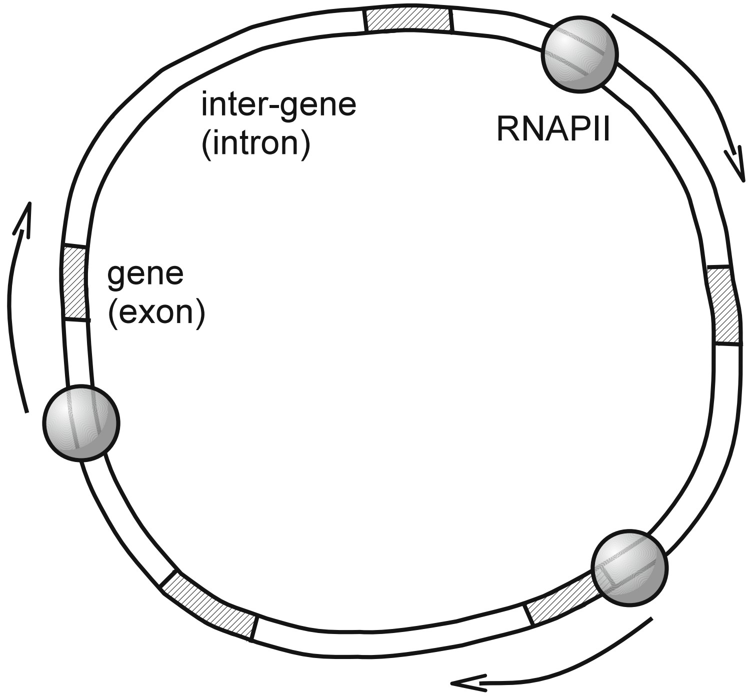


Fig. 1.12 Picture of the traffic flow model with staying time. Exon regions and intron regions appear alternatively and the staying time in exon regions is longer than that in intron ones

assumption, the balls can form only the free flow, and by setting $\tau_j^k - \tau_j^{k-1} = T$ at site j, we have $\tau_{j+1}^k - \tau_{j+1}^{k-1} = T$.

If two adjacent balls are both in an exon region, one should hold $\tau_{j+1}^{k-1} < \tau_j^k$ because the latter ball can move to the next site after the former one leaves there. Then, the $k-1$-th ball first spends the stay time, and k-th one can be adjacent, but the $k-1$-th one moves and creates a vacant space, while k-th one waits for its own waiting time. Then, it never happens that k-th ball cannot move because of $k-1$-th one. That means that there is only free flow in the exon regions. Especially, in the case that k-th ball moves the site immediately after $k-1$-th one leaves, that is, $\tau_{j+1}^{k-1} + 1 = \tau_j^k$ holds, we have $\tau_j^k - \tau_j^{k-1} = \gamma_e$. In the case that $k-1$-th is in an intron and k-th is in an exon, $k-1$-th moves first, and the site becomes vacant. Then, it forms a free flow. The jam can occur only in the case $k-1$-th is in an exon and k-th is in an intron. The time interval preserves if that in the first intron region $\tau_0^k - \tau_0^{k-1}$ is more than γ_e. If not, two balls collide and final time interval becomes γ_e and the k-th ball extra waiting time $\gamma_e = (\tau_0^k - \tau_0^{k-1})$ reduces the time interval of k-th and $k+1$-th ball $\tau_0^{k+1} - \tau_0^k$ to $\tau_0^{k+1} - \tau_0^k - \gamma_e + (\tau_0^k - \tau_0^{k-1}) = \tau_0^{k+1} - \tau_0^{k-1} - \gamma_e$. For such a case, we can determine that k-th and $k+1$-th balls collide with the signature of $\tau_0^{k+1} - \tau_0^k - (\tau_0^k - \tau_0^{k-1} - \gamma_e)$ instead of $\tau_0^{k+1} - \tau_0^k$. Generally, the condition that

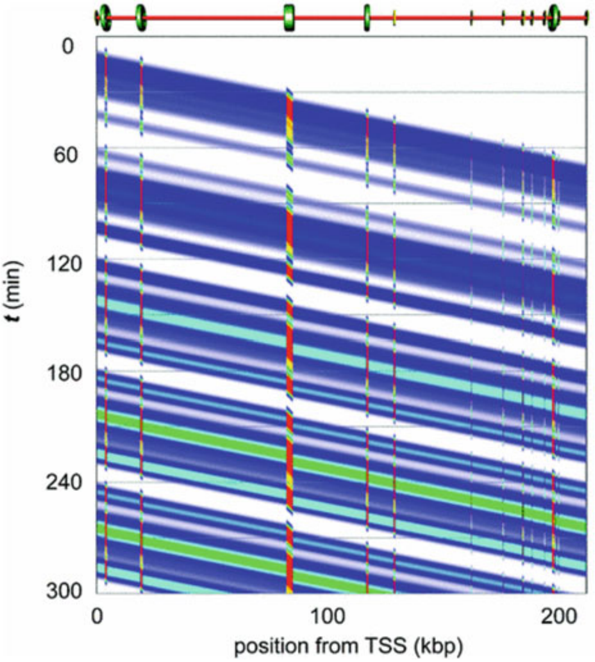


Fig. 1.13 Numerical simulation result of Ohta's cellular automaton model [27]. Colors express the density of RNAPII. Blue regions mean that they are not clouded, and red ones mean that they are clouded

the latter m balls get involved with traffics jams because k-th one overstays at the intron–exon boundary is written in $\tau_0^{k+m} - \tau_0^k - (m-1)\gamma_e$.

Due to the discussion above, if the number of particles is sufficiently small for the number of sites since the traffic jam can occur only in the boundary from the first intron and first exon and the balls after passing this boundary form the free flow, the traffic jam is finally solved.

We can apply similar discussion when there are three types of staying times. The bottleneck appears in the site with the maximal staying sites, and there are no factors that make jams after there.

In [27], the authors introduce periodic boundary condition because it is experimentally known that the transcription start site and the end site are spatially close at the active transcription regions and the RNAPII that finished transcription immediately attaches to the start site and starts transcription again. Figure 1.13 is the numerical simulation results, which match that discussed above because the target gene is sufficiently long.

Finally, we explain a model including RNAPII distribution gaps at introns. As described before, it cannot be explained only by the velocity difference between exons and introns. We have considered the system on the one-dimensional DNA track. However, since the real DNA is in the three-dimensional space and the track

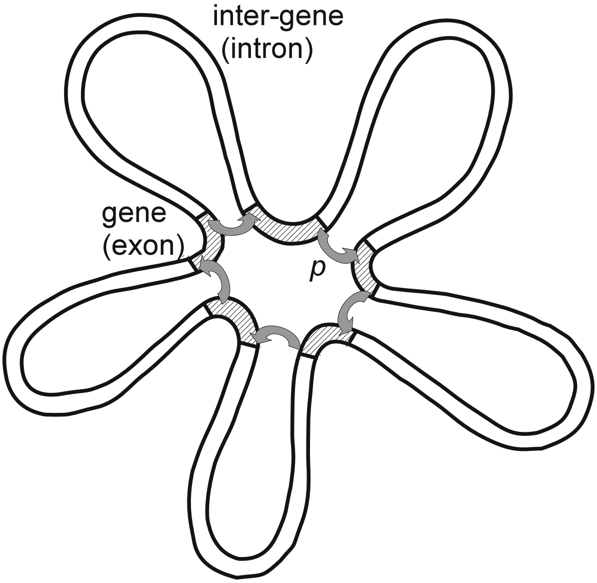


Fig. 1.14 Concept picture of the path-preference cellular automaton model. By being bound by CTCF, some regions of DNA form a ring shape. In such regions, transcriptionally active parts of the DNA, such as exons, are further bound to be spatially close to each other, and RNAPII is free to move through them by spatial diffusion effects

can bend as necessary, it can be possible that two distant sites on the DNA track are spatially close. Therefore, for such sites, one can consider that RNAPII shortcuts by virtue of the diffusion effects or through protein complexes [1].

The authors also proposed a model that enables RNAPII to jump from the end of an exon to the top of the next exon in each gene (Fig. 1.14) [28]. This effect can explain the splicing effects well because transcription products in skipped regions are not degraded but are ever not generated. Furthermore, by changing the target of jumping to the top of other exon regions, one can naturally explain the splicing variants.

The model is basically the same as that proposed before; that is, we consider the traffic flow cellular automaton model with finite amounts of balls and sites and the periodic boundary conditions. Each site is exon or intron. Due to the boundary condition, the number of exonic and intronic regions is the same, and we define the number K. We denote the top and the end of k-th exon as ι_{k-1} and ϵ_k, respectively (we consider the index by modulo K). Without loss of generality, we can set $\iota_K = 1$. We also assume that no more balls are injected after time evolution started in this system.

The basic time evolution rule of the balls is the same as ECA Rule 184, and we do not consider the velocity difference due to exon or intron regions and the proteins binding DNA sites. That is, in all sites, a ball moves to the next site if it is vacant.

Next, we also introduce the jump effect to this model. The ball at $j = \epsilon_k$ moves to $j = \iota_k$ with probability p and moves to the next site $j = \epsilon_k + 1$ by the ECA Rule 184 if the jump fails (probability $1 - p$ or there is another ball at $j = \iota_k$) and stays if there is another ball at the next site in addition to the jump failure. In general, we must first set the ordering to the destinations in the case that several balls are moving to the same site. In this model, we employ the prior for the jumping.

The authors also numerically simulated this model and plot distributions of RNAPII position after sufficient time passed and confirmed the distant peaks. They conclude that there is a diffusion effect because of DNA spatial closeness. Such closeness by RNAPII complexes is already proposed as the name of transcription factory [1, 5, 6, 15, 22, 41, 44].

Let us consider the simplest model with 1-exon and intron and $p = 1$ (that is, the ball at $j = \epsilon_k$ can jump to $j = \iota_k$ if it is vacant). In this model, we can observe a very interesting phenomenon. By calculating the fundamental diagram in the intron region, one finds a non-continuous gap as the number of balls is increasing, and this phenomenon is weakened by strengthening the stochastic effect. The path-preference model is a special case of traffic flow models with bifurcation and confluence. For such models, it is known that such gaps appear in the fundamental diagram [2]. In the path-preference model, we can obtain the exact flow after sufficient time passed by watching the trajectory [24].

In the exon region, only one pair of continuous vacant sites (we call this a *notch*) exists, and balls and empty sites appear alternately in other sites. Several clusters of balls and empty sites appear alternately exist in the intronic region, and other sites are vacant. In the case that there are one cluster and one notch, the system behaves time periodically and the flow is exactly obtained by

$$\langle J_{\mathrm{s}} \rangle = \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} \frac{1}{N_{\mathrm{s}} - 1} \sum_{j=N_{\mathrm{m}}+1}^{N-1} f(j, j+1) = \frac{Q}{T} = \frac{2M - N_{\mathrm{m}} + 1}{2(2M+1)}. \quad (1.4)$$

However, if the length of the cluster M is less than $N_{\mathrm{m}}/2$, the notch arrives at the end of exon before the top of the cluster arrives at the end of the intron. And the notch goes around in the exon again, which changes the periodic orbit pattern. Generally, the condition where the notch goes around exon γ times before the top of the cluster arrives at the end of the intron is written in

$$\frac{N_{\mathrm{s}} - (\gamma - 1)N_{\mathrm{m}}}{2} \leq M < \frac{N_{\mathrm{s}} - (\gamma - 2)N_{\mathrm{m}}}{2}, \quad (1.5)$$

and the flow is expressed as $\langle J \rangle = (2M - N_m + 1)/2/(2M + (\gamma - 1)N_m + 1)$. Surprisingly, under the condition that the number of balls satisfies, the flow takes the same values even if the details of the dynamics behave different. For example, if $\gamma = 2$, the time-space pattern of the system behaves like Fig. 1.15, in which there are one notch and two clusters. Therefore, we can conclude that there exist one notch and two clusters if the number of balls satisfies (1.5) because one of the

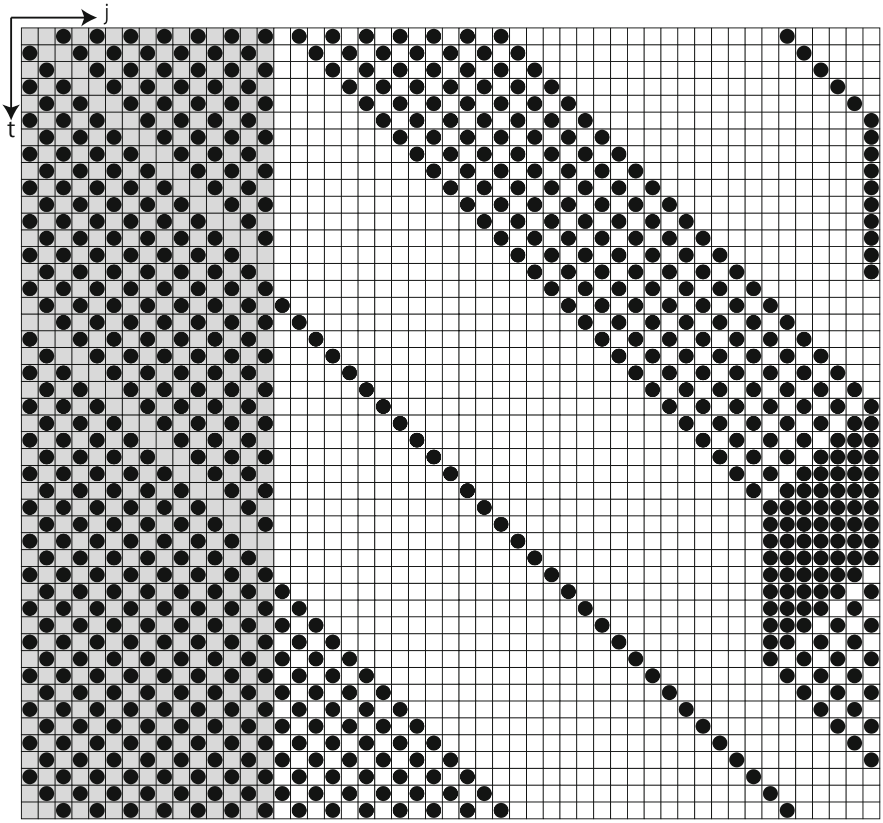


Fig. 1.15 A space-time pattern of the path-preference cellular automaton model. Two clusters of balls travel on the intronic region

clusters has no balls. Anyway, by observing a time-space pattern of periodic orbits, we can calculate the exact value of the flow.

We confirmed that there could be more complicated states with several notches with the same number of balls by setting initial states properly. Such states are established on delicate balances. Then they easily break down with little perturbation such as stochastic effects and reduce to the simpler states explained above.

In this section, we have explained the dynamics of the RNAPII cellular automaton model suggested by Ohta et al. Here, we note that there are other cellular automaton models to explain the dynamics of RNAPII, for example, [3, 4, 13, 26, 38, 39].

Recently, it is believed that the chromatin structure plays an important role in transcription and detecting chromatin structure to understand the transcription mechanics is actively studied. However, it is very difficult to directly observe the chromatin structure, which is the blob of the DNA chain and folding proteins.

The original purpose of authors' model is to explain to construct a model that explains the spatial proximity of distant sites on DNA coordinates during transcription through the movement of RNAPII with a cellular automaton method. However, it has also been found that this model is insufficient to grasp the spatial structure itself directly. One should adopt a more direct approach to capture a dynamical chromatin structure.

One of the major technologies to capture the chromatin structure is to hybridize a specific DNA site with fluorescence labeled probes (3D-FISH method). This approach can observe the structure directly and apply to living cells but only obtain spatial positions of some specific (hybridized) sites. Another method is to aggregate proteins with DNA fragments that contribute to the spatial connection of distant chromatin sites and to detect binding DNA sites from the sequences of the fragments [11]. This method can obtain the connection data of the whole genome, but the data indicate only the adjacency of two chromatin sites and the average of millions of cells due to the experimental method requirement. Therefore, one has to propose a physical or statistical model to guess the chromatin structure from the data [7] and then evaluates whether the estimated structure is correct by the 3D-FISH method.

References

1. Berg, O.G., Winter, R.B., von Hippel, P.H.: Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry **20**, 6929–6948 (1981)
2. Brankov, J.G., Pesheva, N.C., Bunzarova, N.Z.: One-dimensional traffic flow models: Theory and computer simulations. https://arxiv.org/abs/0803.2625
3. Chowdhury, D., Santen, L., Schadsneider, A.: Statistical physics of vehicular traffic and some related systems. Phys. Rep. **329**, 199–329 (2000)
4. Chowdhury, D., Guttal, V., Nishinari, K., Schadschneider, A.: A cellular-automata model of flow in ant trails: Non-monotonic variation of speed with density. J. Phys. A Math. Gen. **35**, L573 (2002)
5. Cook, P.R.: Principles of Nuclear Structure and Function. Wiley, New York (2001)
6. Darzacq, X., Shav-Tal, Y., de Turris, V., Brody, Y., Shenoy, S.M., Phair, R.D., Singer, R.H.: In vivo dynamics of RNA polymerase II transcription. Nat. Struct. Mol. Biol. **14**, 796–806 (2007)
7. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Y.Shen, Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**, 376–380 (2012)
8. Epshtein, V., Nudler, E.: Cooperation between RNA polymerase molecules in transcription elongation. Science **17**, 801–805 (2003)
9. Epshtein, V., Toulmé, F., Rahmouni, A.R., Borukhov, S., Nudler, E.: Transcription through the roadblocks: the role of RNA polymerase cooperation. EMBO J. **22**, 4719–4727 (2003)
10. Fraser, P., Bickmore, W.: Nuclear organization of the genome and the potential for gene regulation. Nature **447**, 413–417 (2007)
11. Fullwood, M.J., et al.: An oestrogen-receptor-α-bound human chromatin interactome. Nature **462**, 58—64 (2009)
12. Galburt, E.A., Grill, S.W., Wiedmann, A., Lubkowska, L., Choy, J., Nogales, E., Kashlev, M., Bustamante, C.: Structural basis of RNA polymerase II backtracking, arrest and reactivation. Nature **471**, 249–253 (2011)

13. Greulich, P., Garai, A., Nishinari, K., Schadschneider, A., Chowdhury, D.: Intracellular transport by single-headed kinesin KIF1A: effects of single-motor mechanochemistry and steric interactions. Phys. Rev. E **75**, 041905 (2007)
14. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., Young, R.A.: A chromatin landmark and transcription initiation at most promoters in human cells. Cell **130**(1), 77–88 (2007)
15. Halford, S.E., Marko, J.F.: How do site- specific DNA- binding proteins find their targets? Nucl. Acids Res. **32**, 3040–3052 (2004)
16. Harima, Y., Takashima, Y., Ueda, Y., Ohtsuka, T., Kageyama, R.: Accelerating the tempo of the segmentation clock by reducing the number of introns in the hes7 gene. Cell Rep. **3**(1), 1–7 (2013)
17. Jin, J., Bai, L., Johnson, D.S., Fulbright, R.M., Kireeva, M.L., Kashlev, M., Wang, M.D.: Synergistic action of RNA polymerases in overcoming the nucleosomal barrier. Nat. Struct. Mol. Biol. **17**, 745–752 (2010)
18. Kanai, M., Nishinari, K., Tokihiro, T.: Stochastic optimal velocity model and its long-lived metastability. Phys. Rev. E **72**, 035102 (2005)
19. Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., Ahringer, J.: Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. **41**, 376–381 (2009)
20. Kulaeva, O.I., Hsieh, F., Studitsky, V.M.: RNA polymerase complexes cooperate to relieve the nucleosomal barrier and evict histones. PNAS **25**(107), 11325–11330 (2010)
21. Li, B., Carey, M., Workman, J.L.: The role of chromatin during transcription. Cell **128**, 707–719(2007)
22. Li, G.W., Berg, O.G., Elf, J.: Effects of macromolecular crowding and DNA looping on gene regulation kinetics. Nat. Phys. **5**, 294—297 (2008)
23. Maekawa, T., Kim, S., Nakai, D., Makino, C., Takagi, T., Ogura, H., Yamada, K., Chatton, B., Ishii, S.: Social isolation stress induces ATF-7 phosphorylation and impairs silencing of the 5-HT 5B receptor gene. EMBO J. **29**(1), 196–208 (2010)
24. Nakata, Y., Ohta, Y., Ihara, S.: Periodic orbit analysis for the deterministic path-preference traffic flow cellular automaton. Japan J. Indust. Appl. Math. **36**, 25–51 (2019)
25. Nishinari, K., Takahashi, D.: Analytical properties of ultradiscrete Burgers' equation and rule-184 cellular automaton. J. Phys. A **31**, 5439–5450 (1998)
26. Nishinari, K., Okada, Y., Schadschneider, A., Chowdhury, D.: Intracellular transport of single-headed molecular motors KIF1A. Phys. Rev. Lett. **95**, 118101 (2005)
27. Ohta, Y., Kodama, T., Ihara, S.: Cellular-automaton model of the cooperative dynamics of RNA polymerase II during transcription in human cells. Phys. Rev. E **84**, 041922 (2011)
28. Ohta, Y., Nishiyama, A., et al.: Path-preference cellular-automaton model for traffic flow through transit points and its application to the transcription process in human cells. Phys. Rev. E **86**, 021918 (2012)
29. Papantonis, A., Larkin, J.D., Wada, Y., Ohta, Y., Ihara, S., Kodama, T., Cook, P.R.: Active RNA polymerases: mobile or immobile molecular machines? PLoS Biol. **8**, e1000419 (2010)
30. Schadschneider, A., Schreckenberg, M.: Cellular automaton models and traffic flow. J. Phys. A Math. Gen. **26**, L679 (1993)
31. Schmittmann, B., Zia, R.P.K.: Statistical Mechanics of Driven Diffusive Systems, vol. 17. Academic Press, New York (1995)
32. Schutz, G.M.: Exactly Solvable models for Many-Body Systems Far from Equilibrium, vol. 19, chap. 1. Academic Press, New York (2001)
33. Schwartz, S., Meshorer, E., Ast, G.: Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. **16**, 990–995 (2009)
34. Shearwin, K.E., Callen, B.P., Egan, J.B.: Transcriptional interference – a crash course. Trends Genet. **21**, 339–345 (2005)
35. Sneppen, K., Dodd, I.B., Shearwin, K.E., Palmer, A.C., Schubert, R.A., Callen, B.P., BarryE-gan, J.: A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli. J. Mol. Biol. **346**, 399–409 (2005)
36. Sutherland, H., Buckmore, W.A.: Transcription factories: gene expression in unions? Nat. Rev. Genet. **10**, 457–466 (2009)

37. Tokihiro, T., Takahashi, D., Matsukidaira, J., Satsuma, J.: From soliton equations to integrable systems through limiting procedure. Phys. Rev. Lett. **76**, 3247 (1996)
38. Tripathi, T., Chowdhury, D.: Interacting RNA polymerase motors on a DNA track: effects of traffic congestion and intrinsic noise on RNA synthesis. Phys. Rev. E **77**, 011921 (2008)
39. Tripathi, T., Schülz, G.M., Chowdhury, D.: RNA polymerase motors: dwell time distribution, velocity and dynamical phases. J. Stat. Mech. P08018 (2009)
40. Wada, Y., Ohta, Y., S.I., et al.: A wave of nascent transcription on activated human genes. Proc. Natl. Acad. Sci. USA **106**(43), 18357–18361 (2009)
41. Wang, F., Greene, E.C.: Single-molecule studies of transcription: from one RNA polymerase at a time to the gene expression profile of a cell. J. Mol. Biol. **412**, 814–831 (2011)
42. Wolfram, S.: Theory and Applications of Cellular Automata. World Scientific Press, Singapore (1986)
43. Wolfram, S.: Cellular Automata and Complexity. Addison-Wesley, Boston (1994)
44. Yao, J., Ardehali, M.B., Fecko, C.J., Webb, W.W., Lis, J.T.: Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation. Mol. Cell. **28**, 978–990 (2007)