Universal Statistics of Cells with Recursive Production

Kunihiko Kaneko[1] [2] Chikara Furusawa[3] [2] ,

[1] Department of Pure and Applied Sciences, Univ. of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
[2] ERATO Complex Systems Biology Project, JST, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
[3] Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

1 Question to be addressed

A cell consists of several replicating molecules that mutually help the synthesis and keep some synchronization for replication. At least a membrane that partly separates a cell from the outside has to be synthesized, keeping some degree of synchronization with the replication of other internal chemicals. Both catalysts and resource chemicals exist in some balance to maintain recursive production. How is such efficient recursive production possible while keeping diversity of chemicals? Is there some universal statistics in abundances of chemicals and in the network for a cell with steady recursive growth?

In a cell, a variety of chemicals form a complex reaction network to synthesize themselves. Then how such cell with a huge number of components and complex reaction network can sustain reproduction, keeping similar chemical compositions?

On the other hand, the total number of molecules in a cell is limited. If there is a huge number of chemical species that catalyze each other, the number of some molecules species may go to zero. Then molecules that are catalyzed by them no longer are synthesized. Then, other molecules that are catalyzed by them cannot be synthesized. In this manner, the chemical compositions may vary drastically, and the cell may lose reproduction activity.

Of course, the cell state is not constant, and a cell may not divide for ever. Still, a cell state is sustained to some degree to keep producing similar offspring cells. We call such condition for reproduction of cell as 'recursive production' or 'recursiveness'. The question we address here is if there are some conditions on distribution of chemicals or structure of reaction network.

Rhe number of each molecule in a cell changes in times through reaction, and the number, on the average is increased for the cell replication. Positive feedback process underlying the replication process may lead to large fluctuations in the molecule numbers. With such large fluctuations and complexity in the reaction network, how is recursive production of cells sustained?[1]

2 Logic

Here, we study statistical characteristics that a cell with recursive growth has to satisfy. In a cell, there is a huge number of chemicals that catalyze each other and form a complex

[1] This type of problem has been discussed in relationship with the origin of life[1, 2, 3]. Here we discuss this problem in connection with a universal feature of a cell with recursive production, in general, which should hold also for the present cell.

network. Through membrane, some chemicals flow in, which are successively transformed to other chemicals through this catalytic reaction network. For a cell to grow recursively, a set of chemicals has to be synthesized for the next generation. Each chemical, for its synthesis, requires some other chemicals as a catalyst. Then, generally speaking, there should exist some mutual relationship among molecules to catalyze each other.

While the number of molecule species is huge in a cell, the number of each molecule species is not necessarily large. Then fluctuations in molecule number are inevitable, since the reaction occurs through stochastic collision. The number of some molecules may go to zero, which sometimes may be dangerous, since such molecules may be essential as a catalyst to the synthesis of some molecules. Then, it is not trivial how recursive production of molecules in a cell is sustained.

Here it will be too much demanding to request that all the molecules keep their number through recursive production of a cell. The compositions of chemicals can differ by cells through divisions. Such loose reproduction should be all right as a beginning of cell. Still, with this loose reproduction, the catalytic activities should be sustained to keep reproduction of cells. In this sense we need to understand how an ensemble of molecules keep catalytic activity and loose reproduction in the amidst of large fluctuations in the chemical compositions.

The biochemical reactions for metabolism, synthesis of membrane and nucleic acid keep some synchrony. In a cell, some transported nutrients are successively transformed to some other chemicals that include catalysts for other reactions. If the transportation of nutrients is higher, that would be helpful for the growth. But if nutrients are too much there will be no room for catalysts from it, and the reaction no longer works. Various chemicals should exist in order for catalytic reactions to work efficiently. On the other hand, if all chemicals exist in the same order, first, the probability of each molecule to meet its catalysts will be lower, and the effective transformation of nutrients is not possible. Furthermore, in this equally distributed number of chemicals, since there are many molecule species that are low in concentration, the reaction events progress just randomly, and the chemical compositions will differ much by generations.

Thus, in order for effective transformation of nutrients, existence of some structure in abundances in chemicals should be favorable. Indeed, in catalytic reactions, there are successive structure, as catalysts for the reactions to transform nutrients, then catalysts for such catalysts for nutrients, and then catalysts for catalysts for catalysts for nutrients, and this cascade continues. It would then be expected that these levels of catalysts do exist in different levels of abundances. On the other hand, if such unbiased distribution in abundances of chemicals exist, the reaction probability is not homogeneous for each reaction. This will lead to decrease the random change of concentrations, as compared with the case of almost equal distribution in numbers.

Hence, for an effective use of resource chemicals, some hierarchical structure in catalysts is favored, with regards to the abundances. Indeed, by taking a specific cell model, we will explicitly show existence of such cascade structure. When a hierarchical structure exists, it is easily expected that some power law exists in the abundances. The distribution $\rho(x)$ that the abundances of chemicals is between x and $x + dx$ is then given by a power law. Indeed, as will be shown, from several model simulations, we find universal power law statistics for a cell that grows efficiently and recursively[4].

At this stage, however, we need to seriously consider the fluctuation of the number of molecules[5]. The chemical abundance of each molecule is under some fluctuation, since the collision process of molecules is basically stochastic. As long as the total number of molecules in a cell is not very huge, fluctuation in each molecule concentration are inevitable. Of course, negative feedback process to stabilize the concentration is one possible solution, to reduce the fluctuations. For reproduction of a cell, however, molecules have to be synthesized which implies some positive feedback process to amplify the number of each molecule species. As for growth, it is better to strengthen this amplification rate, which, however, may amplify the fluctuations also. We need to study some general features of the fluctuations inherent in such positive feedback system. As a very simple illustration, let us consider a process that a moluclue x_m is replicated with the aid of other catalytic molecules.

Then, the growth of the number $N(m)$ of the molecule species x_m is given by $dN(m)/dt = AN(m)$. Here A involves the rate of several reaction processes to synthesize the molecule x_m. Such synthetic reaction process depends on the number of the molecules involved in the catalytic process. Recall, however, that all chemical reaction processes are inevitably accompanied with fluctuations arising from stochastic collision of chemicals. Thus, although the reactions to synthesize a specific chemical and convert it to other chemicals is balanced in a steady state, the fluctuation terms remain. Accordingly, the above rate A has fluctuations $\eta(t)$ around its temporal average $\overline{a}$. Then the above rate equation is written as $dN(m)/dt = N(m)(\overline{a} + \eta(t))$, and it follows that

$$dlogN(m)/dt = \overline{a} + \eta(t). \tag{1}$$

In other words, the logarithm of chemical abundances shows Brownian motion around its mean. The logarithm of chemical abundances is expected to obey normal (Gaussian) distribution, if $\eta(t)$ is approximated by a Gaussian noise, Accordingly, the logarithm of the number molecules is suggested to obey normal (Gaussian) distribution, as

$$P(N) \propto exp(-\frac{(logN - logN_0)^2}{2\sigma}). \tag{2}$$

Such distribution is known as log-normal distribution. In contrast to the Gaussian distribution, the log-normal distribution has a longer tail for more abundances, if plotted in the original scale without taking logarithm.

Of course, the present argument is too simple, since the reaction process is not necessarily directly autocatalytic, as assumed in the above argument. Replication of a molecule occurs only through several steps of reactions. Still, in each term of reaction process, there can appear a multiplicative stochastic process with $\eta(t)x_m$, for the change of x_m, in general. Hence, the log-normal distribution, rather than Gaussian distribution, may be common in a cell that reproduces itself recursively[7, 6]. This argument on this fluctuation is rather primitive, and we need to check if it really works in a model and experiments. We will discuss this problem later.

The log-normal distribution, if plotted in the normal scale, has a very large tail to the abundant side. This large fluctuation in number, in some sense is quite far from the controlled behavior of a cell. Such fluctuations may destroy the recursive growth of a cell. Is there some control mechanism to suppress some fluctuations there? Is gene a controller for such fluctuations? This problem will also be discussed in the last section.

3 Model

What type of a model is best suited for a cell to answer the question raised above? With all the current biochemical knowledge, we can say that one could write down several types of intended models. Due to the complexity of a cell, there is a tendency of building a complicated model in trying to capture the essence of a cell. However, doing so only makes one difficult to extract new concepts, although simulation of the model may produce similar phenomena as those in living cells. Therefore, to avoid such failures, it may be more appropriate to start with a simple model that encompasses only the essential factors

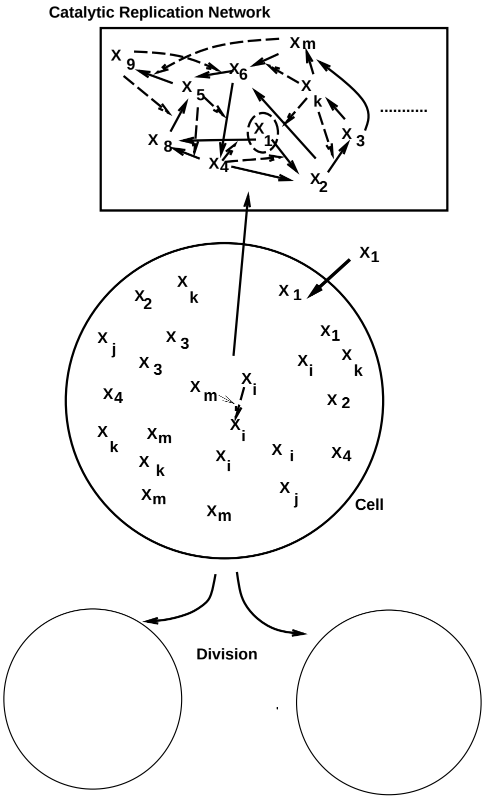

Figure 1: Schematic representation of our modeling strategy of a cell

of living cells. Simple models may not produce all the observed natural phenomena, but are comprehensive enough to bring us new thoughts on the course of events taken in nature.

In setting up a theoretical model here, we do not put many conditions to imitate the life process. Rather we impose the postulates as minimum as possible, and study universal properties in such system. For example, as a minimal condition for a cell, we consider a system consisting of chemicals separated by a membrane. The chemicals are synthesized through catalytic reactions, and accordingly the amount of chemicals increases, including the membrane component. As the volume of this system is larger, the surface tension for the membrane can no longer sustain the system, and it will divide. Under such minimum setup as will be discussed later, we study the condition for the recursive growth of a cell.

Let us start from simple argument for a biochemical process that a cell that grows must at least satisfy. A huge number of chemicals that catalyze each other is spatially arranged in a cell, and in some problems such spatial arrangement is very important, but as a minimal condition, let us discard the spatial configuration of molecules within a cell, since they are not rigidly fixed but can move around randomly to some degree. Hence, we consider just the composition of chemicals in a cell. These molecules change their number through reaction among these molecules. Since most reactions are catalyzed by some other molecules, the reaction dynamics consist of a catalytic reaction network.

Assuming that some reaction processes are fast, they can be adiabatically eliminated. Also, most of fast reversible reactions can be eliminated by assuming that they are already balanced. Then we need to discuss only the concentration (number) of molecules species, that change relatively slowly. For example by assuming that enzyme is synthesized and decomposed fast, the concentrations can be eliminated, to give catalytic reaction network dynamics consisting of the reactions with

$$X_i + X_j \to X_\ell + X_j \tag{3}$$

where X_j catalyzes the reaction[8, 4]. If the catalysis progresses through several steps, this process is replace by

$$X_i + mX_j \to X_\ell + mX_j \tag{4}$$

leading to higher order catalysis[9].

Now, the internal state of the cell can be represented by a set of numbers $(n_1, n_2, \cdots, n_k)$, where n_i is the number of molecules of the chemical species i with i ranging from $i = 1$ to k. For the internal chemical reaction dynamics, we chose a catalytic network among these k chemical species, where each reaction from some chemical i to some other chemical j is assumed to be catalyzed by a third chemical ℓ, i.e. $(i + \ell \to j + \ell)$. The rate of increase of n_j (and decrease of n_i) through this reaction is given by $\epsilon n_i n_\ell / N^2$, where ϵ is the coefficient for the chemical reaction. For simplicity all the reaction coefficients were chosen to be equal, and the connection paths of this catalytic network were chosen randomly such that the probability of any two chemicals i and j to be connected is given by the connection rate ρ.

For a cell to grow, some resource chemicals must be supplied through membrane, which are successively transformed to other chemicals through this catalytic reaction network. These resources (nutrients) are supplied from the environment by diffusion through the membrane (with a diffusion coefficient D). (Note that the nutrient chemicals have no catalytic activity, since they are not products by intra-cellular reactions. Indeed, there does not occur catalytic reactions in the environment.) Besides these nutrients, some of these chemicals may penetrate [2] the membrane and diffuse out while others will not. With the synthesis of the impenetrable chemicals that do not diffuse out, the total number of chemicals $N = \sum_i n_i$ in a cell can increase. As the number of molecules is large enough, the membrane is no longer sustained, even just due to the constraint of surface tension. Then, when the number of molecules is larger than some value, it is expected to divided. We study how this cell growth is sustained by dividing a cell into two when the volume is larger than some threshold. For simplicity the division is assumed to occur when the total number of molecules $N = \sum_i n_i$ in a cell exceeds a given threshold N_{max}. Chosen randomly, the mother cell's molecules are evenly split among the two daughter cells.

Summing up, the basic picture for a simple toy cell we take is given as in Fig.1. In our numerical simulations, we randomly pick up a pair of molecules in a cell, and transform them according to the reaction network. In the same way, diffusion through the membrane is also computed by randomly choosing molecules inside the cell and nutrients in the environment. In the case with $N \gg k$ (i.e. continuous limit), the reaction dynamics is represented by the following rate equation:

$$dn_i/dt = \sum_{j,\ell} Con(j,i,\ell)\ \epsilon\ n_j\ n_\ell/N^2 - \sum_{j',\ell'} Con(i,j',\ell')\ \epsilon\ n_i\ n_{\ell'}/N^2 + D\sigma_i(\overline{n_i}/V - n_i/N), \tag{5}$$

where $Con(i,j,\ell)$ is 1 if there is a reaction $i + \ell \to j + \ell$, and 0 otherwise, whereas σ_i takes 1 if the chemical i is penetrable, and 0 otherwise. The third term describes the transport of chemicals through the membrane, where $\overline{n_i}$ is a constant, representing the number of the i-th chemical species in the environment and V denotes the volume of the environment in units of the initial cell size. The number $\overline{n_i}$ is nonzero only for the nutrient chemicals.

Remark

[2] Even if the reaction coefficient and diffusion coefficient of penetrating chemicals are not identical but distributed, the results reported here are obtained.

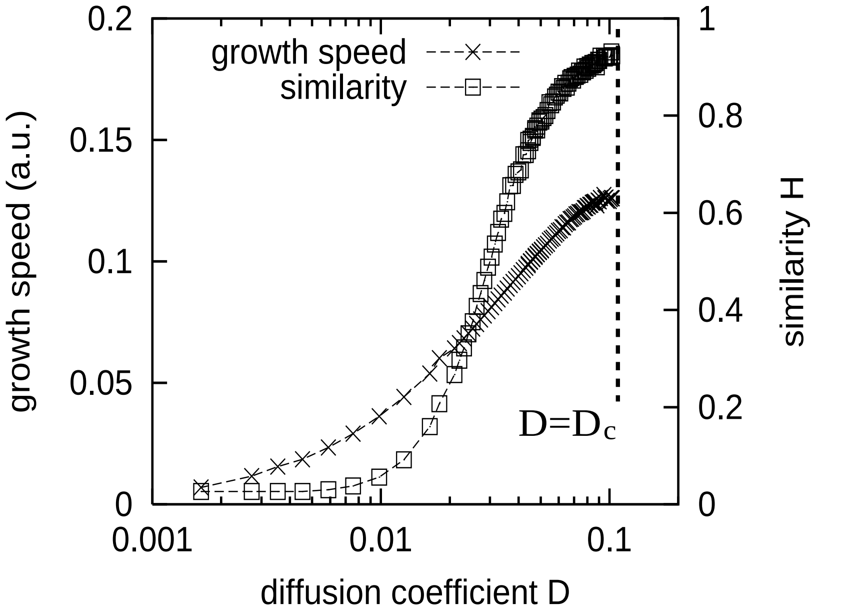

Figure 2: The growth speed of a cell and the similarity between the chemical compositions of the mother and daughter cells, plotted as a function of the diffusion coefficient D. The growth speed is measured as the inverse of the time for a cell to divide. The degree of similarity between two different states m (mother) and d (daughter) is measured as the scalar product of k-dimensional vectors $H(\mathbf{n}_m, \mathbf{n}_d) = (\mathbf{n}_m/|\mathbf{n}_m|) \cdot (\mathbf{n}_d/|\mathbf{n}_d|)$, where $\mathbf{n} = (n_1, n_2, ..., n_k)$ represents the chemical composition of a cell and $|\mathbf{n}|$ is the norm of $\mathbf{n}$. Both the growth speed and the similarity are averaged over 500 cell divisions. Note that the case $H = 1$ indicates an identical chemical composition between the mother and daughter cells. Reproduced from [4].

Models for specific gene expression or signal transduction have been extensively studied these days, which are relevant to understand specific function of a cell. There, we can construct a more detailed model. In our study, we are interested in a recursive production of a 'whole cell', and such partial model is not adequate. Instead, we take a 'crude' catalytic reaction network model that at least include reproduction of the whole set of chemicals.

4 Zipf law

If the total number of molecules N_{max} is larger than the number of chemical species k, the population ratios $\{n_i/N\}$ are generally fixed, since the daughter cells inherit the chemical compositions of their mother cells. For $k > N_{max}$, the population ratios do not settle down and can change from generation to generation. In both cases, depending on the membrane diffusion coefficient D, the characteristics of intra-cellular reaction dynamics change as will be discussed below. Equivalently, this change of intra-cellular dynamics also appear when changing the connection rate ρ.

As D is increased, the growth speed of a cell is increased as shown in Fig.2, since the intake of nutrients is hastened. However, there is a critical value $D = D_c$ beyond which the cell cannot grow continuously. When $D > D_c$, the flow of nutrients from the environment is so fast that the internal reactions transforming them into chemicals sustaining 'metabolism' cannot keep up. In this case all the molecules in the cell will finally be substituted by the nutrient chemicals and the cell stops growing since the nutrients

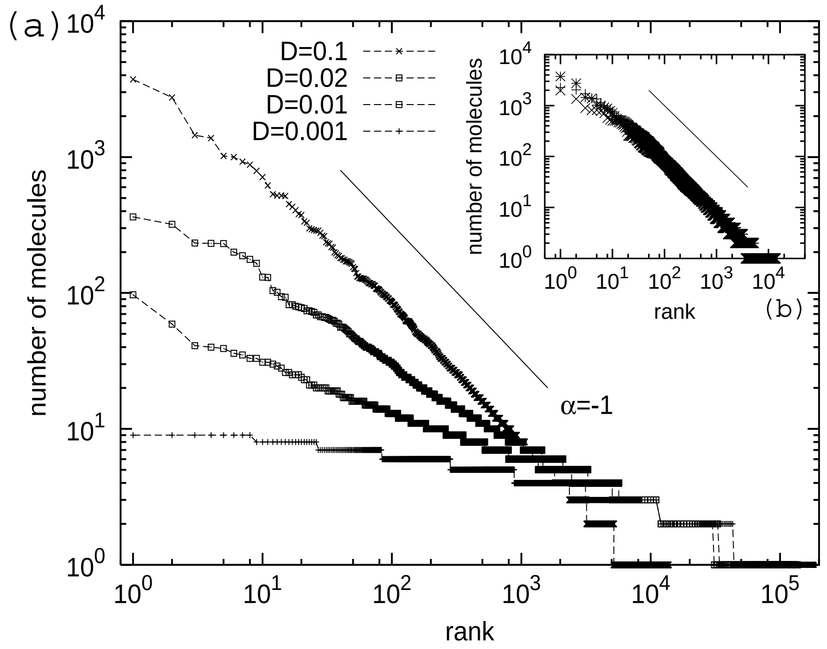

Figure 3: Rank-ordered number distributions of chemical species. (a) Distributions with different diffusion coefficients D are overlaid. The parameters were set as $k = 5 \times 10^6$, $N_{max} = 5 \times 10^5$, and $\rho = 0.022$. 30 % of chemical species are penetrating the membrane, and others are not. Within the penetrable chemicals, 10 chemical species are continuously supplied to the environment, as nutrients. In this figure, the numbers of nutrient chemicals in a cell are not plotted. With these parameters, D_c is approximately 0.1. (b) Distributions at the critical points with different total number of chemicals k are overlaid. The numbers of chemicals were set as $k = 5 \times 10^4$, $k = 5 \times 10^5$, and $k = 5 \times 10^6$, respectively. Other parameters were set the same as those in (a). Reproduced from [4].

alone cannot catalyze any reactions to generate impenetrable chemicals. Continuous cellular growth and successive divisions are possible only for $D \leq D_c$. When the diffusion coefficient D is sufficiently small, the internal reactions progress faster than the flow of nutrients from the environment, and all the existing chemical species have small numbers of approximately the same level. As shown in Fig.2, the growth speed of a cell is maximal at $D = D_c$. This suggests that a cell whose reaction dynamics are in the critical state should be selected by natural selection. Indeed, simulations with evolution of these cells support this[10].

Second, at the critical point, the similarity of chemical compositions between the mother and daughter cell is maximal as shown in Fig.2. Indeed, for $k > N$, the chemical compositions differ significantly from generation to generation when $D \ll D_c$. When $D \approx D_c$, several semi-stable states with distinct chemical compositions appear. Daughter cells in the semi-stable states inherit chemical compositions that are nearly identical to their mother cells over many generations, until fluctuations in molecule numbers induce a transition to another semi-stable state. This means that the most faithful transfer of the information determining a cell's intra-cellular state is at the critical state. To characterize the similarity of cell compositions by divisions, we use the inner product of composition vectors of mother and daughter cells[4, 11, 7]. As plotted in Fig.2, the similarity is maximal around $D \sim D_c$. In this state, chemical compositions of cells are well transferred to offspring through divisions, in the amidst of fluctuations. To sum up, the faithful reproduction is possible for a cell at $D \sim D_c$.

Now, we study the statistics of the abundances of chemicals focusing on the case with $D \sim D_c$. The rank-ordered number distributions of chemical species in our model are plotted in Fig.3, where the ordinate indicates the number of molecules n_i and abscissa shows the rank determined by n_i. As shown in the figure, the slope in the rank-ordered number distribution increases with an increase of the diffusion coefficient D. We found that at the critical point $D = D_c$, the distribution converges to a power-law with an exponent -1. This type of power-law was first studied in linguistics as the frequency of appearance of words by Zipf[12], and is called Zipf law.

The power-law distribution at this critical point is maintained by a hierarchical organization of catalytic reactions, where the synthesis of higher ranking chemicals is catalyzed by lower ranking chemicals. For example, major chemical species are directly synthesized from nutrients and catalyzed by chemicals that are slightly less abundant. The latter chemicals are mostly synthesized from nutrients (or other major chemicals), and catalyzed by chemicals that are much less abundant. In turn these chemicals are catalyzed by chemicals that are even less abundant, and this hierarchy of catalytic reactions continues until it reaches the minor chemical species. In fact, in the case depicted in the inset of Fig.3, a hierarchical organization of catalytic reactions with $5 \sim 6$ layers is observed at the critical point, as schematically shown in Fig.4.

Based on this catalytic hierarchy, the observed exponent -1 can be explained using a mean field approximation. First, we replace the concentration n_i/N of each chemical i, except the nutrient chemicals, by a single average concentration (mean field) x, while the concentrations of nutrient chemicals S is given by the average concentration $S = 1 - k^* x$, where k^* is the number of non-nutrient chemical species. From this mean field equation, we obtain $S = \frac{DS_0}{D+\epsilon\rho}$ with $S_0 = \sum_j \overline{n_j}/V$. With linear stability analysis, the solution with $S \neq 1$ is stable if $D < \frac{\epsilon\rho}{S_0 - 1} \equiv D_c$. Indeed, this critical value does not differ much from numerical observation.

Next, we study how the concentrations of non-nutrient chemicals differentiate. Suppose that chemicals $\{i_0\}$ are synthesized directly from nutrients through catalyzation by chemicals j. As the next step of the mean-field approximation we assume the concentrations of the chemicals $\{i_0\}$ are larger than the others. Now we represent the dynamics by two mean-field concentrations; the concentration of $\{i_0\}$ chemicals, x_0, and the concentration of the others, x_1. The solution with $x_0 \neq x_1$ satisfies $x_0 \approx x_1/\rho$ at the critical point D_c. Since the fraction of the $\{i_0\}$ chemicals among the non-nutrient chemicals is ρ, the relative abundance of the chemicals $\{i_0\}$ is inversely proportional to this fraction. Similarly, one can compute the relative abundances of the chemicals of the next layer synthesized from i_0. At $D \approx D_c$, this hierarchy of the catalytic network is continued. In general a given layer of the hierarchy is defined by the chemicals whose synthesis from the nutrients is catalyzed by the layer one step down in the hierarchy. The abundance of chemical species in a given layer is $1/\rho$ times larger than chemicals in the layer one step down. Then, in the same way as this hierarchical organization of chemicals, the increase of chemical abundances and the decrease of number of chemical species are given by factors of $1/\rho$ and ρ, respectively. This is the reason for the emergence of power-law with an exponent -1 in the rank-ordered distribution. Within a given layer, a further hierarchy exists, which again leads to the Zipf rank distribution.

In general, as the flow of nutrients from the environment increases, the hierarchical catalyzation network pops up from random reaction networks. This hierarchy continues until it covers all chemicals, at $D \to D_c - 0$. Hence, the emergence of a power-law distribution of chemical abundances near the critical point is quite general, and does not rely on the details of our model, such as the network configuration or the kinetic rules of the reactions.

We have checked the universality of the Zipf's law by adopting distributed network connectivity, and distributed parameters, and so forth. Our result is invariant against these changes of the model. This power law in abundances seems to be a quite universal law as long as a cell achieves a recursive growth (i.e., relatively faithful reproduction) and efficient growth. Since the current cell also satisfies relatively faithful reproduction (as well as effective growth for a suitable condition), the above behavior at $D \sim D_c$ may be expected to be true also in the present cell. Then, does this Zipf's law hold in the present cell? In order to investigate possible universal properties of the reaction dynamics, examined are the distributions of the abundances of expressed genes (that are the abundances of the mRNA that produce corresponding proteins). As will be discussed later, the data from several tissues suggest the validity of the present law[4].

remark

Corresponding to the Zipf's law in the rank-abundance relationship, the distribution $p(x)$ of the chemical species with abundance x is proportional to x^{-2}. This is easily obtained by noting that the rank distribution, i.e., the abundances x plotted by rank n can be transformed the density distribution function $p(x)$, which is the probability that the abundance is between x and $x + dx$. Recalling that $dx = dx/dn \times dn$, there are $|dx/dn|^{-1}$ chemical species between x and $x + dx$. Thus, if the abundance-rank relation is given by a power-law with exponent -1, $p(x) = |dx/dn|^{-1} \propto n^2 \propto x^{-2}$.

5 Log-normal distribution

In the lase subsection, we discussed the average abundances of each chemical, and studied the universal statistic over molecule species. Since the chemical reaction process is stochastic, each number of molecule differs by cells. For example, if x_3 in the last section is 3000, it can be 2631, 3203, 2903, 3611,... for each cell. Now we study the distribution of each molecule number, sampled over cells. In Fig.5, we plotted the number distributions of several chemicals in the condition $D \sim D_c$. We measure the number of chemicals when a cell is divided into two. Since the total number of chemicals is constant when the cell is divided, the abundances of each chemical are just proportional to the concentration of

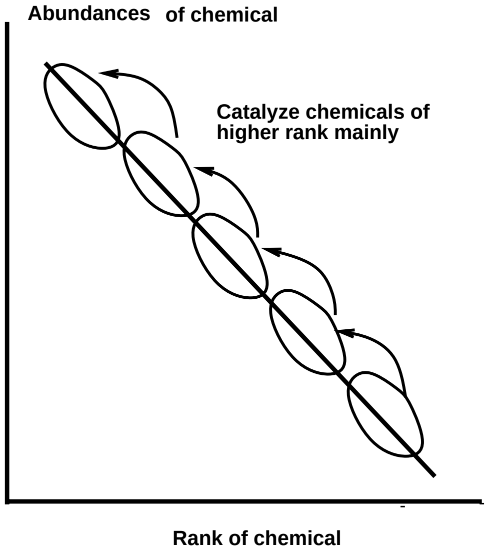

Figure 4: Schematic representation of catalytic cascade. Higher rank chemicals are mainly catalyzed by a lower level. This is a schematic representation showing a rough structure and indeed there are several other cascade paths also.

each. As shown in Fig.5, the distribution is fitted quite well by the log-normal distribution i.e.,

$$P(n_i) \approx exp(-\frac{(logn_i - log\overline{n_i})^2}{2\sigma}), \tag{6}$$

where $\overline{n_i}$ indicates the average of n_i over cells[6].

This log-normal distribution holds for the abundances of all chemicals, except for a few chemicals that are supplied externally to a cell as nutrients, which obey the standard Gaussian distribution. In other words, those molecules that are reproduced in a cell obey the log-normal distributions, while those that are just transported from the outside of a cell follow the normal distributions.

Why does the log-normal distribution law generally hold, in spite of the threat by the central limit due to addition of several fluctuation terms? We have discussed already the possible mechanism for the log-normal distribution in §2. However, there is slight difference here. In the discussion of §2, there is autocatalytic reaction process, as given by $dx_i/dt = cx_jx_i$, which leads to multiplicative stochastic process. In the present example, the reaction process is given by $dx_i/dt = cx_jx_\ell$. Thus, the discussion of §2 is not directly applied.

Still, there is a multiplicative reaction process here. Furthermore, as discussion in §3, there is a cascade reaction process, to support the recursive production around the critical state $D \sim D_c$. There, only a part of possible reaction pathways are used dominantly, to organize a cascade of catalytic reactions so that a chemical in the i-th group is catalyzed by the $(i+1) - th$, and that in the $(i+1) - th$ group is catalyzed by the $(i+2) - th$. A "modular structure" with groups of successive catalytic reactions is self-organized in the network. Then the fluctuations influences multiplicatively through the cascade. By taking logarithm of concentrations (i.e., $logx_m$), these successive multiplication are transformed to successive addition. Now, consider of the average of many random numbers. The distribution of the average approaches the Gaussian distribution, as given by the central limit theorem. The distribution of $logx_m$ is then expected to obey the Gaussian

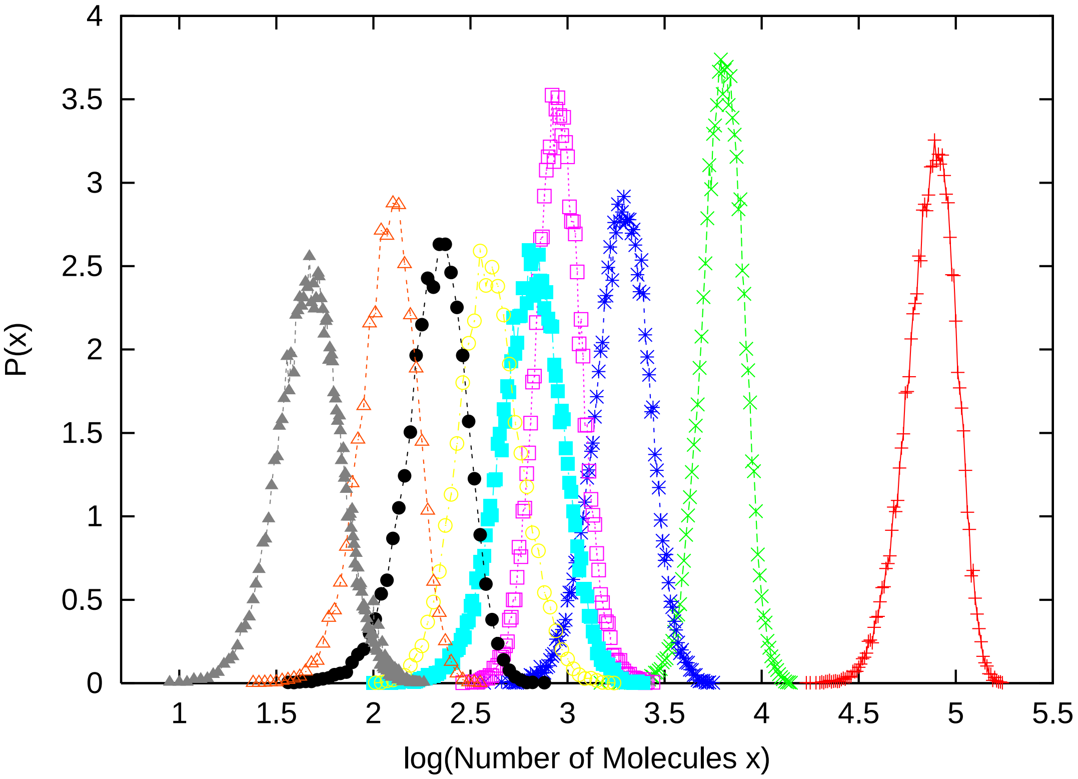

Figure 5: The number distribution of the molecules of chemical abundances. Distributions are plotted for several chemical species with different average molecule numbers. The data were obtained by observing 178800 cell divisions. The parameters were set as $k = 5 \times 10^3$, $N_{max} = 10^6$, and the connectivity to $\rho = 0.02$. 30 % of the chemical species are penetrating the membrane, and the others are not. Within the penetrable chemicals, two chemical species are continuously supplied from the environment as nutrients. The diffusion coefficient D of the membrane was set as 0.04, which is close to the critical value D_c.

distribution. Hence, the log-normal distribution of x_m is derived. Note that, at the critical state we are concerned, this cascade of catalytic reaction continues for all chemical species that are reproduced, and the log-normal distribution holds clearly.

Next we discuss how the magnitude of fluctuations depend on each chemical species. As shown in Fig.5, the width of the distribution does not change much independently of its average abundances, when plotted in the logarithmic scale. This suggests a relationship between the fluctuation and the average for all chemicals. We have thus plotted the standard deviation of each chemical $\sqrt{\overline{(n_i - \overline{n_i})^2}}$ as a function of the average. As shown in Fig.2, the standard deviation (*not* the variance) increases linearly with the average.

To discuss the relationship between the mean and the standard-deviation, one should recall the steady growth and cascade structure in catalytic reactions. Consider two chemicals i and j, one of which (j) belongs to a group of catalyzing molecules for the other. Then the balance between the synthesis and conversion implies $x_i \times A - x_j \times B = 0$, where A and B are average concentrations of other chemicals involved in the catalytic reaction. Assuming the steady growth of a cell, the average concentration satisfies $\overline{x_i}/\overline{x_j} = A/B$. If we further assume the steady state condition for the fluctuations also, then $< (\delta x_i)^2 > / < (\delta x_j)^2 >= (A/B)^2 = \overline{x_i}^2/\overline{x_j}^2$. Hence with this rough argument, the variance is expected to be proportional to the square of the mean, leading to the linear relationship between the mean and standard deviation.

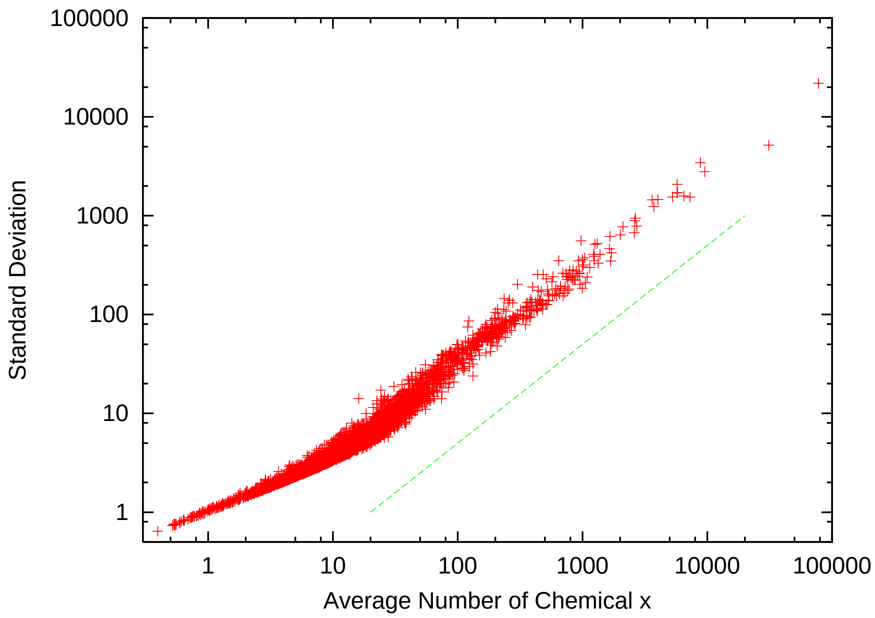

Figure 6: Standard deviation versus average number of molecules. Using the same data set and parameters as for Fig 6.4, the relationship between the average and standard deviation is plotted for all chemical species. The solid line is for reference.

The linear relationship is also found with regards to the variation of chemical abundances by the change of external conditions. For example, we have computed the change from $\overline{x_i}$ to $\overline{x_i'}$ by changing the concentrations of supplied nutrients. The variation $|\overline{x_i'}$ -$\overline{x_i}|$ is again found to be proportional to $\overline{x_i}$ for each chemical i, similarly as the data plotted in Fig.6.

The discovered laws on the distribution and the liner relationship between the average and fluctuation are universally observed, near the critical point with the largest reproduction speed, hold generally and do not rely on the details of the model, such as the network configuration of the kinetic rules of the reactions, as has been confirmed from simulations of a variety of models.

Note, however, that the above arguments for the two laws are based on the steady growth of a cell with catalytic cascade process, realized at $D \sim D_c$. Indeed, as the parameter D is much smaller, all possible reaction pathways are used with a similar weight, where the cascades of catalytic reactions are replaced by random reaction network. In this case, the fluctuations of each molecule number are highly suppressed, and the distribution is close to normal Gaussian. The variance (not the standard deviation) increases linearly with the average concentrations. In other words, the behavior is "normal" as expected from the central limit theorem.

6 Experiment

6.1 confirmation of Zipf's law

First we discuss the validity of Zipf's law with regards to the abundances of many chemicals within the cell. In order to investigate possible universal properties of the reaction dynamics, we examined the distributions of the abundances of expressed genes, that are nothing but the abundances of the mRNA that produce corresponding proteins in a variety of organisms and a variety of tissues. In [4], they used the data publicly available from SAGE (Serial Analysis of Gene Expression) databases [14] over 6 organisms and more than 40 tissues. S AGE allows the number of copies of any given mRNA to be quantitatively evaluated by determining the abundances of the short sequence tags which uniquely identify it [15].

Following the numerical results of the model, we plotted the rank-ordered frequency distributions of the expressed genes, where the ordinate indicates the frequency of the observed sequence tags (i.e. the population ratio of the corresponding mRNA to the total mRNA), and the abscissa shows the rank determined from this frequency. As shown, the distributions follow a power-law with an exponent close to -1 (Zipf's law). We observed this power-law distribution for all the available samples, including 18 human normal tissues, human cancer tissues, mouse embryonic stem cells, nematode (*C. elegans*), and yeast (*S.cerevisiae*) cells. All the data over 40 samples (except for 2 plant data) show the power-law distributions with the exponent in the range from $-1 \sim -0.9$. Even though there are some factors which may bias the results of the SAGE experiments, it seems rather unlikely that the distribution is an artifact of the experimental procedure. Indeed there are increasing supports for this Zipf's law, by using a standard micro array.

6.1.1 Confirmation of laws on fluctuations

Now we discuss the confirmation of the laws of fluctuations. Recalling that the laws are expected to hold for the abundances of a protein synthesized within cells with recursive (steady) growth, we measured the distribution of the protein abundances in *Escherichia coli* that are in the log phase growth, i.e., in a stage of steady growth. To obtain the distribution of the protein abundances, we introduced the fluorescent proteins with appropriate promoters into the cells, and measured the fluorescence intensity by flow cytometry. To demonstrate the universality of the laws, we have carried out several sets of experiments by using a variety of promoters and also by changing places that the reporter genes are introduced (i.e., on the plasmid and on the genome).

Indeed, we have measured the distributions of the emitted fluorescence intensity from *Escherichia coli* cells with the reporter plasmids containing either (enhanced) green fluorescent protein. In general, the fluorescence intensity (the abundance of the protein) increases with the cell size. To avoid the effect of variation of cell size, which may also obey log-normal distribution, we normalized the fluorescence intensity by the volume of each cell, that is estimated by the forward-scatter (FS) signal from the flow cytometry. The distributions of this normalized fluorescence intensity are fitted well by log-normal, rather than Gaussian, distributions, even though each of the expressions is controlled by a different condition of the promoter. We have also measured the abundances of fluorescent protein expressed from the chromosome, which again is found to obey the log-normal distribution. It is furthermore interesting to note that the abundances of the fluorescent proteins, reported in the literature so far, are often plotted with a logarithmic scale[13].

It should be noted that the log-normal distribution of protein abundances is observed when the *Escherichia coli* are in the log phase of growth, i.e., when the bacteria are in steady growth. For other phases of growth, without steady growth, the distribution is found to be often deviated from the log-normal distribution. Note that the theory also supports the log-normal distribution for the steady growth case only, i.e., for a state with recursive production. If a cell is not in a stationary growth state but in a transient process switching from one steady state to another, the universal statistics can be violated.

7 Discussion

In the present chapter we have observed ubiquity of log-normal distribution, in several models. The fluctuations in such distribution are generally very large. This is in contrast to our naive impression that a process in a cell system must be well controlled.

Then, is there some relevance of such large fluctuations to biology? Quite recently, we have extended the idea of fluctuation-dissipation theorem in statistical physics to evolution, and proposed a linear relationship (or high correlation) between (genetic) evolution speed and (phenotypic) fluctuations. This proposition turns out to be supported by experimental data on the evolution of E Coli to enhance the fluorescence in its proteins[16]. Furthermore this phenotypic fluctuation is shown to be tightly related with the genetic variance, measured through phenotype[18]. Hence the fluctuations are important biologically.

The log-normal distribution is also rather universal in the present cell, as demonstrated in the distribution of some proteins, measured by the degree of fluorescence. We have to be cautious here, since too universal laws may not be so relevant to biological function. In fact, chemicals that obey the log-normal distribution may have too large fluctuations to control some function. For example, the abundances of DNA should be deviated from the log-normal distribution. Some other mechanism to suppress the fluctuation may work in a cell[3] .

Note that in a system consisting of molecules synthesized with mutual catalysis, there can appear some key-stone chemicals. These key-stone molecules work as a controlling part. As discussed in [7], this key-stone molecule has a higher connection in reaction network. Then the increase of number fluctuations of such molecule result in a drastic change in the number of other molecules. Accordingly, to have stable recursive production of a cell, such key-stone molecules must have relatively smaller fluctuations, in contrast to the large fluctuations of other molecules arisen from the log-normal distribution. With a combination of such key-stone chemicals, information on recursive growth is generated, that works as a controller for synchronized growth.

Then, what mechanism is a candidate to decrease the fluctuation leading to deviation from log-normal distribution? In relationship with the above key-stone molecules, we previously proposed a minority control mechanism to suppress the fluctuation[19]. In a reproducing cell consisting of mutually catalytic molecules, those in minority have tendency to control the behavior of the cell and are preserved relatively well. These minority molecules are expected to play the role of key-stone molecule species.

The next standard mechanism is negative feedback process. In general, the negative feedback can suppress the response as well as the fluctuation. Still, it is not a trivial question how chemical reaction can give rise to suppression of fluctuation, since to realize the negative feedback in chemical reaction, production of some molecules is necessary, which may further add fluctuations.

The other possible mechanism is the use of multiple parallel reaction paths. If several processes work sequentially, the fluctuations would generally be increased. When reaction processes work in parallel for some species, the population change of such molecule is influenced by several fluctuation terms added in parallel. If a synthesis (or decomposition) of some chemical species is a result of the average of these processes working in parallel, the fluctuation around this average can be decreased by the law of large numbers. Suppression of fluctuation by multiple parallel paths may be a strategy adopted in a cell.

In fact, the minority molecule species in a core catalytic network discussed in [7] has higher reaction paths and has relatively lower fluctuations. This is also consistent with the scenario that more and more molecules are related with the minority species through evolution, as discussed in [19]. With the increase of the paths connected with the minority molecules, the fluctuation of minority molecules is reduced, which further reinforces the minority control mechanism. Hence the increase of the reaction paths connected with the

[3] It is interesting to note that the weight distribution of adult human obeys the log-normal distribution, while the height distribution obeys the Gaussian distribution.

minority molecule species through evolution, decrease of the fluctuation in the population of minority molecules, and enhancement of minority control reinforce each other. With this regards, search for molecules that deviate from log-normal distribution should be important, in future. Here it is important to measure the distribution of chemicals in relationship with its characteristics (such as the connectivity) in the reaction network.

In physics, we are often interested in some quantities that deviate from Gaussian (normal) distribution, since the deviation is exceptional. Indeed, in physics, search for power-law distribution or log-normal distributions has been popular over a few decades, because they are exceptional. On the other hand, a biological unit can grow and reproduce, to increase the number. For such system, the components within have to be synthesized, so that amplification process is common. Then, the fluctuation is also amplified. In such system, the power-law or log-normal distributions are quite common, as already discussed here, and as is also shown in several models and experiments [4, 6]. In this case, the Gaussian (normal) distribution is not so common. Then exceptional molecules that obey the normal distribution with regards to their concentration may be more important.

Note that our findings are not restricted to a cell. As long as a system grows, through production by mutual catalytic processes, universal properties we discussed here will generally appear. Society of human beings with economic production process is such an example, and indeed Zipf law was first studied in human activities[12]. For example, such power law is also observed in the wealth distribution. Universality in a reproduction system revealed in cell biology will be relevant to social dynamics. Indeed, formation of hierarchy, minority control, as well as differentiation that was discussed in reproducing cell models[4, 8, 9, 19] provide conceptual tools to understand history, the dynamics of human society[21]. Furthermore, industry production process, a key issue in the present proceedings, is also such an example of growth system with catalytic process[22]. As discussed in the present volume, the industrial production system is also in the amidst of large fluctuations, and suppression or control of such fluctuations should be important. Ubiquity of log-normal distributions, as well as control mechanisms discussed here may be relevant to such problems.

References

[1] M. Eigen and P. Schuster, *The Hypercycle* (Springer, 1979).

[2] F. Dyson, *Origins of Life*, Cambridge Univ. Press., 1985

[3] S.A. Kauffman, *The Origin of Order*, Oxford Univ. Press. 1993

[4] C. Furusawa and K. Kaneko, Phys. Rev. Lett. 90 (2003) 088102

[5] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002) Science 297, 1183

[6] C. Furusawa, T. Suzuki, A. Kashiwagi, T. Yomo and K. Kaneko ; Ubiquity of Log-normal Distribution in gene expression, BIOPHYSICS, 2005, in press.

[7] K. Kaneko, *Phys. Rev. E* **68** 031909 (2003)

[8] K. Kaneko and T. Yomo, B. Math.Biol. 59 (1997) 139 J. Theor. Biol., 199 243-256 (1999)

[9] Furusawa C. & Kaneko K., Bull.Math.Biol. 60; 659-687 (1998); Phys Rev Lett. 84:6130-6133 J. Theor. Biol. 209 (2001) 395-416; Anatomical Record, 268 (2002) 327-342

[10] C. Furusawa and K. Kaneko, "Evolutionary origin of power-laws in Biochemical Reaction Network; embedding abundance distribution into topology", submitted to Phys. Rev. Lett.

[11] D. Segré, D. Ben-Eli, D. Lancet, Proc. Natl. Acad. Sci. USA 97 (2000)4112; D. Segré et al., J. theor. Biol. **213** (2001) 481

[12] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949).

[13] W. J. Blake, et al. *Nature* **422**, 633 (2003)

[14] A.E. Lash et al.,*Genome Research* **10**(7), 1051 (2000). V.E. Velculescu et al., *Cell* **88**, 243 (1997): S. J. Jones et al., *Genome Res.* **11**(8), 1346 (2001):

[15] V.E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler, *Science* **270**, 484 (1995).

[16] K. Sato, Y. Ito, T. Yomo, and K. Kaneko; Proc. Nat. Acad. Sci. USA 100 (2003) 14086-14090

[17] Ito, Y., Kawama, T., Urabe, I. & Yomo, T., J. Mol. Evol. 58 (2004) 196

[18] K. Kaneko and C. Furusawa, "An Evolutionary Relationship between Genetic Variation and Phenotypic Fluctuation", submitted to PNAS.

[19] Kaneko K, Yomo T. J. Theor. Biol. 312 (2002) 563-576

[20] Matsuura T., Yomo T., Yamaguchi M, Shibuya N., Ko-Mitamura E.P., Shima Y., and Urabe I. Proc. Nat. Acad. Sci. USA 99 (2002) 7514-7517

[21] K. Kaneko and A. Yasutomi, "Historical Science in view of Complex Systems Studies", in *Current Status of Economics*, ed. M. Yoashida, in Japanese, 2005

[22] A. Ponzi, A. Yasutomi and K. Kaneko, "Economic Cycles in an Evolving Economic Production Network", JEBO, in press