氏　　名	小笠原　理
学位（専攻分野）	博士（理学）
学位記番号	総研大乙第147号
学位授与の日付	平成１７年９月３０日
学位授与の要件	学位規則第６条第２項該当
学位論文題目	Statistical Analysis of Anatomical Expression Pattern and Expression Strength
論文審査委員	主　査　教授　五條堀　孝 教授　舘野　義男 教授　西川　建 教授　城石　俊彦 所長　加藤　菊也（大阪府立成人病センター）

論文内容の要旨

The advent of whole-genome sequencing and large-scale profiling of gene expression has revealed several unexpected phenomena in the genome that had never been discovered from the studies of a small number of genes. With the examination of the factors responsible for the formation of such new phenomena, it is conceivable that new relationship between underlying biological processes will be elucidated. In this thesis, I examine such phenomena in the transcriptome of a large variety of organisms, using data from public databases and those obtained in our laboratory and I report unexpected relationships in the transcriptome evolution.

Zipf's law of transcriptome is one of the phenomena that have been revealed by genome-wide studies of gene expression. This law states that there is a relationship between the transcript frequency (f) and abundance rank (r) represented as $f=k/r^b$, where k is a constant and b is a constant parameter that represents the absolute value of the slope in a log-log plot of transcriptome frequencies. I reported in my published paper that this law was applicable to all human normal tissues that I observed. Further, in muscle and liver, which are primarily composed of a homogeneous population of differentiated cells, the slope parameter b was nearly equal to 1. In cell lines, epithelial tissue and compiled transcriptome data, only high-rankers deviate from the law. In addition to my work, several other authors also reported this law in other species. It has been known that this law is applicable to a large variety of species, such as vertebrates (*Homo sapiens, Mus musculus* and *Rattus norvegicus*), invertebrates (*Drosophila melanogaster* and *Caenorhabditis elegans*), other eukaryotes (*Saccharomyces cerevisiae* and *Arabidopsis thaliana*) and even to bacteria (*Escherichia coli*). It is remarkable that the observed slope parameter b is almost unique ($b\approx1$) irrespective of the species investigated.

To explain the factors responsible for the formation of the law, I proposed an evolutionary

model of Zipf's law of transcriptome. In this model, Zipf's law could be replicated based on three assumptions. The first assumption states that the baseline expression level of each gene in a cell is coded in the genome sequence such as in the *cis*-elements or enhancer regions; therefore, the expression levels of genes are affected by mutations, and these are inherited to the offspring. This assumption is supported by the fact that in a large variety of organisms, the expression levels of genes have abundant natural variation and familial aggregation. In addition, it is known that the location of the *cis*-element and/or the *trans*-factor of the genes can be determined with the quantitative trait locus (QTL) analysis in which the expression level of each gene is treated as a quantitative trait. The second assumption states that the expression level changes in stochastic proportion to its intensity. This assumption is supported by the observation that expression differences accumulate at a constant ratio in primates and rodents. The third assumption states that the number of expressed genes in a cell type is nearly constant throughout the evolutionary process and that any functional gene is prohibited from losing its gene expression ability. By the Monte-Carlo simulation of the model, I showed that a stable distribution of f=0.1/r was obtained from these three assumptions, regardless of the initial distribution. To demonstrate that the uniqueness of slope parameter b among variety of species can be replicated from the evolutionary model, I conducted a Monte-Carlo simulation to determine the condition for converging the distribution with the slope parameter $b \approx 1$. In my model, the slope parameter b depends on the number of mRNA molecules in a cell (M), the number of genes expressed in the cell (G), and the permissible lower limit of expression level (L). When the value of parameter M is fixed to 300,000, as in the case of a typical human cell, and L is set to a sufficiently small value, i.e., 1–2 copies/cell, the distribution converged to $b \approx 1$ over a wide range of values of parameter G, i.e., from 10,000 to 50,000 genes in a cell. This is the reason for the universality of b, which is predicted by the evolutionary model of Zipf's law.

At approximately the same time, several authors (Kuzunetosov 2003, Frusawa and Kaneko

2003, Ueda 2004) proposed other models of Zipf's law of transcriptome independently (see reference in the thesis 20, 21, 22). All of their models attributed the Zipf's law of transcriptome to the gene expression dynamics in each cell (the dynamics model, hereafter). They argued that the change in gene expression level in each cell follows the formulation of geometrical Brownian movement. However, in addition to the lack of reliable biological evidence for the dynamics model, I pointed out that the dynamics model cannot replicate the observed Zipf's law of transciptome even in mathematical sense, contrary to the authors' assertion. From the dynamics model, it follows that the rank order of expression level in each cell independently diverged at random, even if the same type of cell was considered. It is noteworthy that the observations of expression level distribution were obtained from a mixture of millions of cells. The central limit theorem of probability theory states that the distribution obtained from such a mixture of a large number of completely diverged samples should be a normal distribution. Therefore, if the dynamics model is valid, the observed distribution of the expression level of genes should follow a normal distribution, not a Zipf's law distribution. Such divergence does not occur in my evolutionary model; hence, Zipf's law is replicated.

Obviously, the determination of the correct cause of Zipf's law of transcriptome critically influences the direction of further investigation. Based on the dynamics model of Zipf's law, Ochiai et al. (2004) derived a formula that describes the elementary process of gene expression dynamics in a cell. Determining such a formula is crucial for estimating gene regulatory networks from time course data of gene expression profiles. However, if my assertion is valid, the proposed formula will lose its ground. If my evolutionary model is accepted, Zipf's law of transcriptome would be related to the neutral model of transcriptome evolution, proposed by Khaitovich et al. (2004) (53). They discovered a clocklike accumulation of gene expression divergence within primates and rodents (53). These results were in agreement with the observation of Rifkin et al. (2003), who reported that differences in gene expression were consistent with phylogenetic relationships among

Drosophila species(12). Zipf's law of transcriptome can be viewed as a new support for the clocklike accumulation of expression diversity, because the assumption of my evolutionary model is nearly equivalent to the neutral model of transcriptome evolution.

In the evolutionary model of Zipf's law, I focused only on the evolutionary change in expression levels of genes in a cell. Next, I tried to extend my study to the expression patterns in various tissues (anatomical expression pattern, hereafter). To investigate the evolution of anatomical gene expression patterns, I focused on housekeeping genes as a special set of genes that were definitely expressed and function in all cell types.
Identification of housekeeping genes from large-scale expression profiles was first exemplified by Velculescu et al. (1999) who used the SAGE method(24). This was followed by Warrington (2000) and Hsiao (2001) who used oligonucleotide microarrays (25, 26). These studies opened up new opportunities to explore the relationship between expression patterns and other features of genes, such as gene length, sequence divergence, location in the chromosomes, and so on. Several sets of housekeeping genes were published along with such studies, but it has rarely been well discussed whether or not the analyzed set of genes is a non-biased representative of housekeeping genes. In fact, by comparing the two published screenings for housekeeping genes, one based on the GeneChip method and the other based on the SAGE method, I found that there was a low concordance between the results of the two screening methods. I also found that, in both processes, there was poor sensitivity in the identification of housekeeping genes. Therefore, I examined the causes of this inconsistency, and by tuning the parameters for housekeeping gene selection, I compiled a more reliable set of housekeeping genes. In this study, I found a good correlation between the observed breadth of gene expression (the number of organs in which gene expression was detected) and the expression level of genes, even in a set of known housekeeping genes. Based on this, I concluded that the expression level of a gene seriously affects the apparent breadth of its expression. This was particularly manifested in the

result where I succeeded in doubling the number of housekeeping gene candidates (from 2,792 to 5,537) without losing specificity. The newly identified housekeeping genes (new HK) and the previously identified housekeeping genes (old HK) shared features in terms of constancy of expression abundance among tissues (expression evenness), cellular localization of products, and the fraction of genes that have CpG islands at their transcription start sites. Estimated contaminants, which comprise approximately 12%–20% of either new or old HK, were genes that were unique to widely distributed cells rather than those that were common to a wide variety of cells.

Main points of this thesis are summarized as follows.

Part I

1. I reported that mRNA frequencies in human normal tissues obeyed the Zipf's law. Especially, in the organs that are primarily composed of a homogeneous population of differentiated cells, the slope parameter was nearly equal to 1.
2. I proposed a new theoretical model for explanation of the factors responsible for the formation of the Zipf's law. It is the evolutionary model of the Zipf's law of transcriptome. Further, I gave several experimental supports for the each assumption of the model.
3. I concluded that the gene expression dynamics models for the Zipf's law of transcriptome are not valid because they can not replicate the Zipf's law even in mathematical sense.

Part II

In order to extend my study from the gene expression strength in a tissue type to the expression patterns in various tissues (anatomical expression pattern), I focused on housekeeping genes as a special set of genes that were definitely expressed and function in all cell types.

1. By comparing the two representative large scale screenings for housekeeping genes in

the human genome, I found that there was a low concordance between the results of the two screening methods and there was poor sensitivity in the identification of housekeeping genes.

2. I demonstrated that the cause of the low concordance was that the expression level of a gene seriously affects the apparent expression breadth (the number of tissues in which the gene was expressed), because there was a good correlation between the observed breadth of gene expression and the expression level of genes, even in a set of known housekeeping genes.

I compiled a new and more reliable set of housekeeping genes, and I succeeded in doubling the number of housekeeping gene candidates (from 2,792 to 5,537) without losing specificity.

論文の審査結果の要旨

小笠原理氏より提出された論文「Statistical Analysis of Anatomical Expression Pattern and Expression Strength」は、Part I と Part II の２つに分かれている。

Part I では、生体において転写産物の量(f)とそれが発現する組織や器官における大小をランク(r)としてみたとき、$f=k/r^b$（kは定数）という特別な関係式が成り立つということが提唱されており、これを Zipf's law というが、これに申請者は注目した。申請者は、トランスクリプトームにおいて、遺伝子発現量が決まる分子機構やその進化メカニズムを解明する目的で、この Zipf's law が人体の遺伝子発現で成立するかどうかを調べた。申請者が人体の様々な組織や器官でゲノムワイドな調査を行った結果、その Zipf's law が人体で十分に成立するだけでなく、b というパラメタがほぼ１に近いことも発見した。また、マウス等の他の生物でも同様の解析を行った結果、他の生物でもこの b の値がほぼ１に近いことがわかった。これらの発見を説明するため、申請者は独自のトランスクリプトームの進化モデルを考案し、コンピュータシミュレーションによって b が１に近い条件を調べたところ、１細胞あたりの mRNA 分子の数や発現する遺伝子の数が現在観察されている値に近いものであればよいことを突き止め、この申請者の進化モデルが今までの結果と両立することを明らかにした。さらに、今まで提案されているいくつかのトランスクリプトームの力学モデルでは、Zipf's law を再現的に支持できないことから、この申請者の浄化淘汰に基づく進化モデルが現在最もトランスクリプトームの観察結果を説明できるモデルであると考えられる。また、申請者は、この進化モデルは、遺伝子発現量の生物種間の進化的な分岐過程における差異が分岐時間に比例するというトランスクリプトームの中立進化説も説明できることを明らかにした。

PartII では、PartI で用いたトランスクリプトームのデータやその他の遺伝子チップや SAGE データを用いて、いわゆるハウスキーピング遺伝子の遺伝子発現を詳細に調べた。その結果、いままで知られているハウスキーピング遺伝子は約 2,800 個程度であったが、実際には約 5,500 個以上も存在する可能性のあることを明らかにした。これらのハウスキーピング遺伝子の数の推定は、PartI で行われた遺伝子発現の進化モデルをさらに精緻にするなど、今後の発展に重要な知見を与えるものである。

申請者は、2003 年７月より約２年間にわたって国立遺伝学研究所、生命情報・DDBJ 研究センター遺伝子発現解析研究室でプロジェクト研究員として大久保公策教授のもとに「遺伝子発現の統計的解析とその生物学的意義」についての研究に従事しており、その研究成果が本申請論文の主なる内容である。

当該研究を中心とした発表原著論文が６編あり、そのうち申請者を筆頭著者として国際誌へ１本が既に発表され、そのほかにもう１本は投稿中である。

このように、申請者の研究は、遺伝子発現に関して新しい展開をもたらすとともに生物情報解析における先導的研究と位置づけられ、優れた業績に基づいていることから、博士論文審査委員会は、申請論文を論文博士の学位を授与するに値するとの結論に達した。