Chapter 28
Scaling Behavior of Maximal Repeat Distributions in Genomic Sequences: A Randomize Test Follow Up Study

J.D. Wang
Asia University, Taiwan
Ka-Lok Ng
Asia University, Taiwan

ABSTRACT
The maximal repeat distribution analysis is redone by plotting the relative frequency of maximal repeat patterns against the rank of the appearance. In almost all of the cases, the rank plots give a better coefficient of determination values than the authors’ previous work, i.e. frequency plot was used. A randomized version is repeated for the maximal repeat study; it is found that rank plot regression analysis did not support scaling behavior; hence, the validity of the findings is not due to an artifact.

INTRODUCTION
In our previous study (Wang et al., 2008) it is found that the relative frequency distribution of maximal repeat sequences $\log(P(k))$ versus the frequency of appearance $\log(k)$ exhibits scaling behavior, i.e. $P(k) \sim k^{-\gamma}$. In order to validate the above findings are not due to an artifact, we redone the study by performing the randomize test, and reanalysis the results by plotting the relative frequency of maximal repeat patterns against the rank of the appearance. The rank plots give a better coefficient of determination values than our previous work, i.e. using frequency plot. The rank plot study is motivated by Li et al. (Li et al. 2005). The work gave a rigorous presentation on the theory of scale-free graphs, and it is pointed out that size-rank plot is more reliable than size-frequency plot.

In section 2, we give a description of the randomize test procedures. Section 3 is the power-law study results. Section 4 is the summary and discussion.

DOI: 10.4018/978-1-60566-902-1.ch028
METHOD

The randomize test proceed as the following; (i) the nucleotides positions are randomized while keeping the nucleotide composition fixed, that is the percentage of nucleotides A, T, C and G in the randomize sequence is the same as the original sequence, and (ii) the maximum repeat distribution analysis is repeated five times via the external memory approach (Wang 2006), the average scaling exponent and coefficient of determination are computed.

RESULTS

In this section, we present the maximal repeats frequencies study in different genomic sequences. To implement our analysis using the power-law model, we first record the total number of identified maximal repeats ($N$) that are appear in the genomic sequence under study. For each frequency of appearance of maximal repeat, $a$, we record the total number ($N(a)$) of maximal repeat patterns that have such a frequency of appearance in that genomic sequence. The frequency of appearance of maximal repeat $a$ is ranked in descending order, for example, the frequency of appearance of maximal repeat two ($a=2$) is ranked number one, that is $k$ equals to one. We divide number $N(a)$ by $N$ and call it $P(k)$, and then plot $\log P(k)$ against $\log(k)$, where $k$ denotes the rank of $a$ with non-zero $P(k)$. Power law states that $P(k) \sim k^{-\gamma}$. $P(k)$ is the fraction of the total number of maximal repeats with rank $k$ ($1 \leq k \leq 999$) in the genomic sequence under study, for example, $P(902)$ is the fraction of maximal repeat patterns among the total number of maximal repeat patterns that appears in the genomic sequence one thousand times, and 1000 is ranked 902. The length of the maximal repeat pattern we search ranges from 3 to 50 bps.

Table 1 is the results of the regression analysis of the genomic sequences chosen from different groups of taxa. For each of the genomic sequences, we give the result for the total genomic sequence length, exponent of the power-law and $r^2$. GenBank ID column label with ‘chromosomes’ means the power-law result obtained with the species’ chromosomes length assembled together. Regression analysis determines that the exponent $\gamma$ ranges from 1.81 to 2.06 (the same range as our previous work [1]), and all the 56 species’ data are well fitted by the power-law distributions (the date reject the null hypothesis of uniform distribution with $p$-value $<<10^{-6}$). All the 56 species have a $r^2$ value larger than 0.938.

Table 2 is the results of the regression analysis of plant chloroplast genome sequences. Regression analysis determines that $\gamma$ lies between 1.79 and 1.88, and all the 17 species’ data are well fitted by the power-law distributions ($r^2 > 0.930$).

Results of the maximal repeat distributions regression analysis of plant mitochondria genome sequences is the same as our previous work (Wang et al. 2008) reported, so it is omitted.

In order to valid the above findings are not due to an artifact, twelve genome sequences are picked (ten from nucleus, one from chloroplast, and one from mitochondria) to perform the randomize test, and the results are reported in Table 3. It is found that randomized studies do not provide consistent results for the frequency-rank plot regression analysis, that is, some sequences give poor $r^2$ values while reasonable values are obtained in other cases. For instance, the average exponent of the randomized *E.coli* K-12 sequence is 2.13 with an average $r^2$ value of 0.490, whereas the yeast chromosome IV has an average exponent of 1.99 with an average $r^2$ value of 0.840.
Related Content

Logical Connections of Statements at the Ontological Level
Cungen Cao, Yuefei Sui and Yu Sun (2010). *International Journal of Cognitive Informatics and Natural Intelligence* (pp. 59-85).
www.igi-global.com/article/logical-connections-statements-ontological-level/45185?camid=4v1a

A Computational Simulation of the Cognitive Process of Children Knowledge Acquisition and Memory Development
www.igi-global.com/chapter/computational-simulation-cognitive-process-children/72286?camid=4v1a

Contribution and Extensions
www.igi-global.com/chapter/contribution-extensions/27883?camid=4v1a

A Cognitive Informatics Reference Model of Autonomous Agent Systems (AAS)
www.igi-global.com/article/cognitive-informatics-reference-model-autonomous/1578?camid=4v1a