J. Appl. Genet. 43(3), 2002, pp. 391-395

Monte Carlo simulation of genome viability with paralog replacement

Stanisław CEBRAT, Dietrich STAUFFER[1]

Department of Genetics, Institute of Microbiology, University of Wrocław, Wrocław, Poland

Abstract. Recent analyses of genome content have revealed that many single functions, even in haploid organisms, can be executed by more than one gene. As a result, experimental disruption of many individual genes does not exert lethal effects on the organism or even any visible change in the phenotype of the organism with a knockedout gene. Our analysis shows that such genetic redundancy allows for an appreciably higher mutation load in the genome simulations before the viability of the whole organism is destroyed.

Key words: DNA, genome, Monte Carlo simulations, mutation, paralog.

Introduction

The latest progress in whole genome sequencing has revealed redundancy in genetic information. In many genomes, the number of genes which can be knocked out without any visible phenotypic effect is substantial. In the unicellular eukaryote *Saccharomyces cerevisiae* there are about 5350 protein coding genes (MACKIEWICZ et al. 1999, 2002), of which only 924 are essential and probably unique, since their elimination from the genome has a lethal effect, while for about half of the other genes no changes in phenotype after gene disruption have been found (MIPS 2002 DATABASE).

Comparative intra- and intergenomic studies of different coding sequences have shown that many sequences present in the same genome are homologous. These sequences, if occurring in the same genome, are called paralogs. Paralogs

Received: March 1, 2002. Accepted: June 21, 2002.

Correspondence: S. CEBRAT, Institute of Microbiology, Wrocław University, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland, e-mail: cebrat@microb.uni.wroc.pl

[1] Present and permanent address: Institute for Theoretical Physics, Cologne University, D-50923 Köln, Euroland.

can perform the same function, complementing each other, or they can fulfil different functions. Furthermore, a single function can be executed by genes which are not paralogs. SŁONIMSKI et al. (1998) found a specific distribution of paralogs sharing homology. They noticed that the total number of paralogs belonging to the groups with *n* paralogs in each group is twice as high as the number of paralogs belonging to the groups with *n*+1 paralogs in each group. This observation seems to be a universal law for all genomes sequenced thus far. In this paper we try to estimate how the observed redundancy enhances the viability of genomes under the mutation pressure.

Monte Carlo simulations

Computer simulations using random numbers are called Monte Carlo simulations, after the roulette tables of the Mediterranean casino. Physicists usually see METROPOLIS et al. (1953) as the beginning, but Metropolis himself wrote in METROPOLIS (1987) that Enrico Fermi already in the 1930's simulated some random walks without publishing them. We do not need here the more sophisticated methods of importance sampling (LANDAU, BINDER 2000) for thermal equilibrium. To perform a computer instruction with a probability *p* one calculates a random number r between zero and unity, and then follows this instruction if and only if $r < p$. We produced such random numbers by multiplying an initial odd integer again and again by the integer 16807, a simple and commonly used method.

Results and discussion

It would be too simplistic to assume that all *N* genes in any genome are essential for the survival of the organism and are stored in DNA as only one copy. Some genes, like these determining the colour of the hair, are not crucial for survival, and some genes are repeated in the genome, thus resulting in several genes exercising one function. Let us thus assume that an organism has *L* functions essential for survival; of the *L* essential functions, *K* are executed by single copy genes, while the other *L*-*K* essential functions could be performed by multi-copy genes stored in several regions of the genome. Some of them have evolved by duplication of a single copy gene, showing sequence homology, and are called paralogs. In our calculations we have assumed that only such sequences can complement a function and represent the information redundancy in the genomes. If a "healthy" paralog is present it could function properly as a replacement for the mutated gene. Presumably, not many functions are unimportant for simple organisms: *N*-*L* is at most of the order of *L*, and thus perhaps about one quarter of all essential functions are performed by products of the single-copy genes: $L = 4K$.

For diploid organisms, a mutation of one of the K essential and unique genes still allows the individual to survive if the mutation is recessive and in the second DNA sequence (haplotype) the deleterious mutation of the corresponding gene (allele) have been avoided. Thus, we restrict ourselves to the simpler haploid organisms, like bacteria or haploid yeasts. Let $q = 1 - p$ be the probability of an essential gene to be mutated. We estimate the maximum mutation probability q_c or minimal reliability $p_c = 1 - q_c$, which allows the whole organism to survive; more precisely, for $p = p_c(L)$ the probability R of survival reaches 0.5:

$$R(p,L) > 1/2 \text{ for } p > p_c(L) \quad . \tag{1}$$

We now assume the various mutations to be identically and independently distributed, and the same happens to the paralogs. Then trivially, $p^K \geq R \geq p^L$, with $R \approx p^K$ if lots of paralogs are available to replace defect genes, while $R \approx p^L$ if paralogs are rare. According to SŁONIMSKI et al. (1998), SŁONIMSKI (1999), and TIURYN, RADOMSKI, SŁONIMSKI (1999, 2000) the probability that the "original" gene have n paralogs is $\pi_n = 2^{-n-1}$, which is 1/2 for $n = 0$; thus $K=L/2$, which is appreciably higher than the above estimate $K \approx L/4$. This is possible, because many paralogs, after a long evolution, could fulfil functions different than originally. Thus, we allow for additional replacement effects, called pseudo-paralogs, to decrease the number K of irreplaceable essential genes from $L/2$ to $L/4$. These pseudo-paralogs have the same reliability p as the original gene and its paralogs. Paralogs and pseudo-paralogs together are called replacements (group of genes mutually complementing their function) and are added to the one original gene.

If n_i is the number of replacements for function i, with $i = 1, 2, \ldots, L$, then the failure probability for function i is q^{ni+1} and the survival probability R of the whole organism is:

$$R = \prod_i \left(1 - q^{n_i + 1}\right) . \tag{2}$$

The simplest assumption is to give all functions exactly the same number n of possible replacements (paralogs or pseudo-paralogs). Then the survival probability is

$$R = (1 - q^{n+1})^L \cong \exp(-Lq^{n+1}) \tag{3a}$$

for $q^n << 1$, giving for $R = 1/2$:

$$\text{In}(2) / L = q_c^{n+1} \quad \text{or} \quad q_c = (0.7 / L)^{1/(n+1)}. \tag{3b}$$

With $L = 300$ and $n = 1$ the maximally allowed failure rate is thus q_c » 0.05. This seems to be a reasonable number if nature achieves reliability more by redundancy then by reliability of every element (GAVRILOV, GAVRILOVA 2001), while technology uses little redundancy and a much higher reliability (lower q) for the single elements. If $n = 0$ instead for all functions, i.e. without any replacements, $q_c = 0.0023$ is much smaller.

The above estimate, however, is too optimistic since the number n fluctuates from function to function and is distributed so that $K \approx L/4$ of the L essential genes have no replacement, $n = 0$. Taking only paralogs with probability $\pi_n = 2^{-n-1}$, as mentioned above, and thus $L = 2K$, we get from a straightforward Monte Carlo simulation $q_c = 0.045$, 0.0046, 0.0005 for $L = 30$, 300, and 3000, respectively. Adding pseudo-paralogs with a Poisson distribution of average 0.7 we get $L = 4K$ and higher thresholds $q_c = 0.083$, 0.0091, 0.0009.

Table 1. Values of q_c threshold for different distributions of paralogs (see text for details)

Paralog distribution	Threshold q_c for varying genome size [L]		
	30	300	3000
Exponential	0.045	0.0046	0.0005
Poisson	0.083	0.0091	0.0009
Connectivity*	0.039	0.004	0.0004

*Values of q_c threshold calculated using the standard Hoshen-Kopelman algorithm for connectivity for both exponential and Poisson distributions

If instead we require some connectivity in genomic space, we may apply percolation theory (STAUFFER, AHARONY 1994, BUNDE, HAVLIN 1996, SAHIMI 1994) for a square lattice with L horizontal lines of varying length $1 + n_i$ each. Each line i stores the original gene and its possible n_i replacements. The organism then is defined as viable if a path of unmutated genes or replacements on nearest-neighbour lattice sites connects the top line ($i = 1$) with the bottom line ($i = L$). We check for the percolation threshold q_c, defined so that on average half of the organisms survive, i.e. that half of the lattices percolate from top to bottom. Using the standard Hoshen-Kopelman algorithm (STAUFFER, AHARONY 1994) for connectivity, we get with the same exponential distribution of the number of paralogs and Poisson distribution of the number of pseudo-paralogs: $q_c = 0.039$, 0.004, 0.0004, not as good as the simpler criterion above. These distributions are chosen so that again $L = 4K$.

In summary, the introduction of suitable replacements into haploid genomes allows an up to four times higher error rate q_c, like 0.0091 instead of 0.0023.

In further studies we will try to show what mechanisms could generate this specific distribution of paralogs observed by SŁONIMSKI et al. (1998). Our preliminary evolutionary simulation in the spirit of TIURYN et al. (1999), which worked best, was the following algorithm:

Initially all L functions are represented in n =15 sequences (1 original and 14 paralogs). At each iteration each individual first undergoes the Verhulst test, i.e. it dies with probability N_{pop}/N_{max} where N_{pop} is the current size of the population and N_{max} is a parameter often called the carrying capacity. Then each organism only survives with probability $(p_i)^n$, where n is the largest number of paralogs for any of its L functions. For each survivor we then determine the probability R that all L functions are still working even though each paralog works only with probability p. Those who survive produce four offspring and die. At birth, for each of the *L* functions and each paralog, one mutation occurs which with probability 0.01 either increases (with probability 0.4) or decreases (probability 0.6) the number of paralogs by 1.

Further studies of different phenomena responsible for generation of the information redundancy in genomes, particularly those restricting the genome size should be undertaken.

Acknowledgements. We thank Jan RADOMSKI (Interdisciplinary Center for Mathematical and Computational Modeling, Warszawa, Poland) for an encouraging discussion.

REFERENCES

BUNDE A., HAVLIN S. (1996). Fractals and disordered systems. Springer, Berlin-Heidelberg.

GAVRILOV L.A., GAVRILOVA N.S. (2001). The reliability theory of aging and longevity. J. Theor. Biol. 213: 527-545.

LANDAU D.P., BINDER K. (2000). A guide to Monte Carlo simulations in statistical physics. Cambridge University Press, Cambridge UK.

MACKIEWICZ P., KOWALCZUK M., GIERLIK A., DUDEK M.R., CEBRAT S. (1999). Origin and properties of noncoding ORFs in the yeast genome. Nucleic Acid Research 27(17): 3503-3509.

MACKIEWICZ P., KOWALCZUK M., MACKIEWICZ D., NOWICKA A., DUDKIEWICZ M., DUDEK M.R., CEBRAT S. (2002). How many protein coding genes are there in the *Saccharomyces cerevisiae* genome? Yeast 19: 619-629.

METROPOLIS N. (1987). The beginning of the Monte Carlo method. Los Alamos Science 15: 125.

METROPOLIS N., ROSENBLUTH A.W., ROSENBLUTH M.N., TELLER A.M., TELLER E. (1953). Equation of state calculation by fast computing machines. J. Chemical Physics 21: 1087-1092.

MIPS (2002) – http://mips.gsf.de/proj/yeast/.

SAHIMI M. (1994). Applications of percolation theory. Taylor and Francis, London.

SŁONIMSKI P.P., MOSSE M.O., GOLIK P., HENAUT A., DIAZ Y., RISLER J.L., COMET J.P., AUDE J.C., WO NIAK A., GLEMET E., CODANI J.J. (1998). The first laws of genomics. Microbial and Comparative Genomics 3: 46.

SŁOMINSKI P.P. (1999). Comparison of complete genomes: Organization and evolution. Proc. 3rd Annu. Conf. Comput. Molecul. Biol., ACM Press, Stanislaw Ulam Lecture: 310.

STAUFFER D., AHARONY A. (1994). Introduction to percolation theory. Taylor and Francis, London.

TIURYN J., RADOMSKI J.P., SŁONIMSKI P.P. (1999). Striking properties of DNA molecules. Comptes Rendus Acad. Sci. Paris, series III, 322: 455-459.

TIURYN J., RADOMSKI J.P., SŁONIMSKI P.P. (2000). A formal model of genomic DNA multiplication and amplification. In: Comparative Genomics (Sankoff D., Nadeu J.H., eds.). Kluwer, Netherlands: 503-513.