SHORT COMMUNICATION

Gradimir Jankovic · Milorad Pavlovic
Vladimir Lazarevic · Djordje Vukomanovic

Rates of nucleotide substitution, mutation at a locus, and the "beanbag" gene number in man

Received: October 26, 2001 / Accepted: January 16, 2002

Abstract We estimated the number of different human genes by relating the patterns of spontaneous mutation at the population and individual level. A geometric distribution model of mutation was used in which the average rates of nucleotide replacement (P) and mutation at a locus (p), obtained by experiment, were used to determine the estimate of the physical size of the coding genome (n) in man. The probabilistic relation used, $P = (1 - p)^{n-1}p$, integrates two different referential time scales of mutation, that of a nucleotide and year and that of a coding gene and generation. The estimates of n, for different values of P and p, are compatible with the experimentally determined genome sizes. The size of the coding portion of the genome appears to be evolutionarily constrained by an interplay between the rate of nucleotide replacement and the pattern of mutation at the level of the individual locus. The evolution of the size of the coding genome may be more dependent on the number of generations than on time.

Key words Substitution rate · Mutation rate · Coding genome · Genome size · Gene number

G. Jankovic (✉) · V. Lazarevic
Institute of Hematology, University Clinical Center, 11000 Belgrade, ul. Dr. Koste Todorovica br. 2, Yugoslavia (Serbia)
Tel. + 011-361-7777 ext. 3711; Fax + 011-361-5564
e-mail: gradjank@Eunet.yu

M. Pavlovic
Institute of Infective Diseases, School of Medicine, Belgrade, Yugoslavia (Serbia)

D. Vukomanovic
Institute of Mathematics, Serbian Academy of Sciences and Arts, Faculty of Civil Engineering, University of Belgrade, Belgrade, Yugoslavia (Serbia)

Introduction

The public consortium's estimate of a complete set of human genes is 32 000 (International Human Genome Sequencing Consortium 2001). Celera predicts that there are around 39 000 genes, but the evidence for some 12 000 of these is weak (Venter et al. 2001). Recent estimates based on sampling have also placed the number of human genes at 25 000–35 000 (Ewing and Green 2000; Roest Crollius et al. 2000). We envisage a simple relation that integrates the nucleotide substitution rate in the course of evolution, mutation rate at a locus, and number of coding nucleotides in the genome. Although an instance of "beanbag genetics" (Crow 2001), it may give us a quantitative insight into an evolutionary process that maintains the amount of genetic information contained within the genome.

There is a clear distinction between mutant (gene) substitution at the population level and gene mutation at the individual level. The amino acid/nucleotide differences between species reflect the results of mutant substitution, or evolution, rather than simply gene mutations. We are mostly concerned here with mutations that affect a single nucleotide (point mutations) in which one nucleotide (base) is replaced by another. The average rates of mutation at a classical gene locus are estimated to be ~10^{-5} per generation (Mukai and Cockerham 1977; Chakraborty and Neel 1989). They are much higher than the estimates for base substitution rates for two reasons. First, a generation may represent 20–30 cell divisions (2^{20}~10^{6}; therefore, 20 cell divisions can produce 1 million cells). Thus the value of ~10^{-5} suggests a rate per cell division of between 10^{-6} and 10^{-7}. A second reason that per-locus rates are much higher than base acquisition rates is that between 100 and 1000 of the possible substitutions are detectable. This detection rate is a consequence of the fact that synonymous and undetectable mutations that alter the amino acid composition (i.e., are functionally irrelevant) remain silent. The measured rates per conventional locus per replication should then be greater by two or three orders of magnitude — as in fact they are.

Method

Our treatment is conceptually related to an accuracy of replication in which a limit on the size of the genome can be placed for any given replication accuracy. Consider a genome of n nucleotides and the probability that an error is made in replication of u per nucleotide. Then $Q = (1 - u)^n \approx e^{-nu}$, where Q is the probability that a sequence produces an exact copy of itself. Hence, the maintenance of adaptation requires, very roughly, that $nu < 1$.

Metaphorically, assume that the quality control examines successive nucleotide positions along a continuous string of unit nucleotide sequences. Each unit sequence has an average probability p of having a nucleotide mutate. Then the probability, P, that a mutant nucleotide comes up for the first time on the nth successive nucleotide position on the string can be conveniently expressed as $P = (1 - p)^{n-1}p$. For very large n, $P = (1 - p)^n p$ is acceptable. A unit nucleotide sequence can be homologized (Jankovic 1984) to a continuity of a gene's coding sequence (average length, 1340 bp [International Human Genome Sequencing Consortium 2001]) with an average mutation rate, p, of 10^{-7} per gene copy per replication (Kimura 1983; Vickers et al. 2000). Once it is realized that the point mutation comes up for the first time at an nth successive position on the continuous linear array of unit nucleotide sequences, a P could be perceived, under neutrality, as equivalent to the probability of substitution per site, n denoting the entire length of the coding genome. The number of nucleotides up to and including the nth one, which is first in the entire sequence to be affected by mutation, reflects the interplay between the nucleotide replacement rate per year (P) and the mutation rate per sub-sequence/locus per generation (p). Given the average values of P and p, the extent of the coding genome that is spared from potentially hazardous change by mutation (i.e., the total amount of coded message that can be evolutionarily maintained, n) could be derived. That the size of the coding part of the genome may have evolved as a product of the concerted evolutionary balance between P and p, notwithstanding an apparent discrepancy of timescales, may be appreciated once the evolution is viewed as a process operating at two levels of DNA organization: (1) a pervasive differentiation of structure operating at the level of the nucleotide (P), and (2) an integration of function operating at the level of individual genes (p). A limit is consequently placed on the size of the coding genome evolved (n), for any given P and p, such that the coded information maintains the degree of integratedness that guarantees its evolutionary value as a functional whole. Here the whole is defined by the pattern of functional relation between its parts (individual genes), not by the sum of its parts. The probabilistic relation used links mutation on two distinct (but related) referential time frames, implying thereby that the generation time itself may be an independent factor in the evolution of the size of the coding genome.

The human coding genome

Assuming that DNA sequences fix mutations at a selectively neutral rate of roughly 5×10^{-9} substitutions per site per year (Kimura 1983), then the extent of the coding genome follows from $5 \times 10^{-9} = (1 - 10^{-7})^n 10^{-7}$, whence $n \sim 3.0 \times 10^7$ nucleotides or ~22400 genes ($3.0 \times 10^7/1340$).

The preceding argument uses a nucleotide substitution rate in neutral noncoding DNA to estimate the gene number in man. The average autosomal substitution rate of 1.28×10^{-9} per site (Nachman and Crowell 2000) specifies the number of genes as ~32500 ($4.35 \times 10^7/1340$). This estimate reproduces a set of ~32000 genes determined by experiment (International Human Genome Sequencing Consortium 2001). The evolutionary rate of fixing alleles in a population and the mutation rate at a locus may appear to capture an underlying evolutionary rationale for a particular amount of information contained within the human genome (for example, minimizing the cost of mutational compromise of the locus). However, it is not really necessary to invoke any rationale (i.e., selection); the cause of the relationship, from the neutralist interpretation, is purely statistical. The size of the coding genome is an artifact imposed by historical chance that may mimic a rationale, but not purposely.

The viral and mitochondrial genomes

For the large DNA viruses with substitution rates of $\sim 10^{-8}$ per site per year and mutation rates of $\sim 10^{-7}$ per replication, the integral genome sizes predicted on the model are of the order of 10^5 nucleotides, as indeed they are. For the mitochondrial genome comprising ~16000 bp, and at a synonymous rate of mtDNA substitution estimated to be 5.7×10^{-8} per site and year (Brown et al. 1982), the p is predicted to be roughly 5.7×10^{-4} per gene and generation [$5.7 \times 10^{-8} = (1 - p)^{16150}p$]. Although the mutation rate of mitochondrial genes is certainly exceptionally high, the rate of 5.7×10^{-4} is bound to be inaccurate due to deviation from the model as the coding fraction of genome markedly diminishes.

Results and discussion

The present estimate of the absolute size of the protein-coding portion of the human genome, based on the rates of molecular evolution and gene mutation together, is in accordance with the neutralist belief that mutation plays an important role in molecular evolution, so that the rate of molecular evolution is mainly determined by the rate of mutation.

The probabilistic relation used, $P = (1 - p)^{n-1}p$, links mutation on distinct (but related) time scales, implying thereby that the *generation time* may be quite an independent factor in the evolution of the size of the coding genome. This is not entirely unexpected because spontaneous mutation is much more dependent on the number of cell generations than on time (Kuick et al. 1992).

The neutralist view that the rate of nucleotide substitution should be higher for an organism with a higher mutation rate than for an organism with a lower mutation rate is also reinforced by the present study. Absolute sizes of the protein-coding portion of different genomes are consistent with the operation of neutral evolution at the level of global abundance of informative DNA.

Acknowledgments We are very grateful to the anonymous reviewer and would like to acknowledge the insightful comments which were very helpful in the elaboration of the time-frame problem of mutation rates.

References

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Biol 18:225–239

Chakraborty R, Neel JV (1989) Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc Natl Acad Sci USA 86:9407–9411

Crow JF (2001) The beanbag lives on. Nature 409:771

Ewing B, Green P (2000) Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet 25:232–234

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Jankovic GM (1984) The meaning of homology. Nature 310:635

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, pp 175, 195, 314

Kuick RD, Neel JV, Strahler JR, Chu EHY, Bargal R, Fox DA, Hanash SM (1992) Similarity of spontaneous germinal and in vivo somatic cell mutation rates in humans: implications for carcinogenesis and for the role of exogenous factors in “spontaneous” germinal mutagenesis. Proc Natl Acad Sci USA 89:7036–7040

Mukai T, Cockerham CC (1977) Spontaneous mutation rate at enzyme loci in *Drosophila melanogaster*. Proc Natl Acad Sci USA 74:2514–2517

Nachman, MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304

Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F, Saurin W, Weissenbach J (2000) Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet 25:235–238

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. (2001) The sequence of the human genome. Science 291:1304–1351

Vickers M, Brown GC, Cologne JB, Kyoizumi S (2000) Modelling haemopoietic stem cell division by analysis of mutant red cells. Br J Haematol 110:54–62