Genome Evolution by Substitutions, Duplications and Deletions

Yi Zhou and Bhubaneswar Mishra
Biology Department, New York University, 100 Washington Square East, New York, NY 10003, USA;
Courant Institute of Mathematical Sciences, New York University,
251 Mercer Street, New York, NY 10012, USA; and
Watson School of Biological Sciences, Cold Spring Harbor Laboratory,
1 Bungtown Rd., Cold Spring Harbor, NY 11724, USA.
(Dated: July 3, 2003)

Recently, detailed statistical analyses of sequenced genomes have provided support for the "evolution by duplication" theory proposed by S. Ohno. Based on Ohno's theory, we suggest a parsimonious model consisting of substitutions, duplications, and deletions, and estimate the parameters of this model at various scales (word sizes) over several genomes. We conclude that deletions play as critical a role in these models as other evolutionary mechanisms, and therefore the omission of the deletion process leads to an inadequate model. We also present an analysis of the parameters to this model across species, leading to a better understanding of the biological processes that modulate duplications, deletions, and substitutions.

PACS numbers: Valid PACS appear here

The genome of an organism stores its genetic information required for all cellular processes. It is obvious that a better understanding of genome composition, structure and evolution in various organisms is critical for biological studies. Somewhat surprisingly, a deeper understanding of genome statistics also leads to the design of better bioinformatic algorithms and tools, such as genome assembly, probe design, and comparative genomics, etc. — each playing an important role in genomics science. However, not till recently has it become apparent that genomes are neither random nor deliberately and accurately sculpted. The seemingly random non-coding regions have nonrandom compositions and long-range correlations, whereas the more conserved coding regions are subject to constant mutations and tolerant of enough polymorphisms.

Further detailed analyses on the genomic sequences lead to the discovery that some statistical characteristics of genome composition and structure are generic in different organisms, in spite of the huge diversity at the sequence level. For example, all the genomes are characterized by the over-representation of high-frequency components, which are observed as the "fat-tails" in the histograms of mer (oligonucleotides of a certain length) frequencies, gene family sizes, and duplication copy numbers [1][2]. Interestingly, this statistical feature appears also to be reflected in higher-level cellular processes, such as protein-protein interaction networks, metabolic networks, and genetic pathways [3][4]. Those observations are evidences for "evolution by duplication" — a theory for genome evolution originally proposed by S. Ohno in 1970 [5]. The theory suggests that duplication is one of the main driving force in genome evolution. Based on this theory several research groups have proposed genome evolution models that incorporate two basic processes: duplications and substitutions (point mutations). DeLisi *et al.* [6] described a simple model to explain the gene family size distributions in various microbes. Very recently, Lee *et al.* [7] proposed another minimal model that was able to fit the 6-mer (oligonucleotides of length 6) distributions in several bacterial genomes.

Our parsimonious model [1] for genome evolution incorporates not only substitutions and duplications, but also deletions. Based on our analyses on different models, we found that deletions play a role no less critical than substitutions or duplications. The effect of deletion process cannot simply be replaced by a reduction in duplication rate and/or an increase in substitution rate. Those conclusions from model analyses are consistent with biological experimental results [8], which show that deletions happen as often as duplications, and their contribution in shaping the genome composition is significant. Our model, which considers all three processes, is able to fit the distributions of not only 6-mers, but also mers of other sizes from a wide range of scale. It applies equally well to both prokaryotic and eukaryotic genomes.

In our model, a genome is represented by a directed Eulerian multi-graph. Each pair of inverse-complementary mer species of a particular length is represented by a node [12]. Whenever two non-overlapping mers are immediately adjacent to each other in the genome, they are connected by an additional directed edge. Without loss of generality, the edges are always directed from the 5' end to the 3' end. In a graph created in this manner, the number of directed edges from node i to node j ($k_{i,j}$) indicates how many times the i^{th} mer is immediately adjacent to the 5' end of the j^{th} mer in the genome. Due to the Eulerian property of the graph, each node has identical in- and out-degrees. We use k_i to represent both the out-degree (${k_i}^{out}$) and the in-degree (${k_i}^{in}$) of the node i, which are equal to the copy number of the corresponding mer in the genome. For mers of size l, and a genome of length L, the graph will have a total of $N = \frac{4^l}{2}$ nodes and $E = \frac{L}{l} = \sum_{i=1}^{N} k_i$ edges. Each possible Eulerian path in the nontrivial (non-singleton) connected component encodes a genome with the same mer composition.

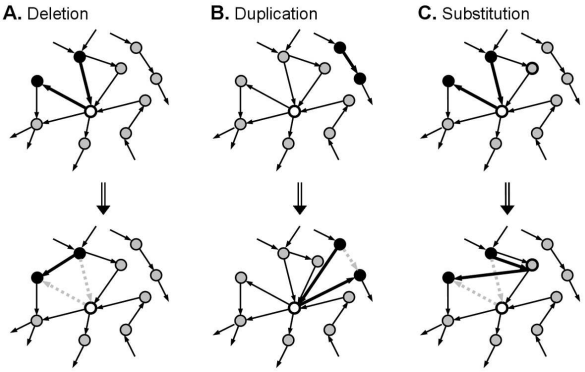

FIG. 1: The three processes occurring during graph evolution: *deletion*, *duplication*, and *substitution*. In each process, the target node (clear circle) is chosen with preference for nodes with larger degrees. In deletion (**A**), a pair of edges of the target node (thick black arrows), one incoming and one outgoing, is randomly chosen and deleted, and a new edge (thick black arrow) is added between the ascendent and descendent nodes (black filled circles). In duplication (**B**), new edges are added between the target node and the ascendent/descendent nodes (black filled circles) of an edge (thick black arrow) randomly chosen to be deleted. In substitution (**C**), a randomly chosen pair of edges of the target node (thick black arrows), one incoming and one outgoing, is rewired to the randomly chosen substitute node (gray filled circle with thick boundary). Note that all the processes during graph evolution preserve the equality of the in-degree and out-degree of each node.

However, the genomes represented by the same graph do not necessarily have the same arrangement of mers.

The evolution of a genome is modeled as a stochastic evolution process on the multi-graph going through multiple iterations. The model assumes that all the presently existing genomes originated from a very small proto-genome with uniformly randomly distributed mers. Thus, the initial graph is a random graph with a small average degree. In each iteration, one of the three possible processes occurs: *duplication* of a chosen mer (with probability p_1), *deletion* of a chosen mer (with probability p_0), or *substitution* of a chosen mer by another mer (with probability q) (Figure 1). Therefore, $p_1+p_0+q=1$.

To avoid extinction, we let $p_1 > p_0$. During graph evolution, let ${k_i}^t$ and E^t indicate the copy number of i^{th} mer and the total number of mers in the evolving genome at t^{th} iteration. If we assume that the target mers for any process is chosen uniformly randomly from the genome, then the probability of i^{th} mer species being chosen for a process in the next iteration is proportional to its frequency in the genome in the current iteration ($\propto \frac{{k_i}^t}{E^t}$). Such a strategy implements a "rich gets richer" dynamic rule, and is reminiscent of Polya's Urn model [9]. If a mer undergoing substitution is modeled as changing into any other mer with equal probability after substitution [13], then with this simplifying assumption, we can write down the difference equation describing the expected probability distribution for the copy number of the i^{th} mer:

$$\begin{aligned} P({k_i}^t = n) &= P({k_i}^{t-1} = n-1)P({k_i}^t = n \mid {k_i}^{t-1} = n-1) + P({k_i}^{t-1} = n)P({k_i}^t = n \mid {k_i}^{t-1} = n) \\ &\quad + P({k_i}^{t-1} = n+1)P({k_i}^t = n \mid {k_i}^{t-1} = n+1) \\ &= P({k_i}^{t-1} = n-1)\left(p_1\frac{n-1}{E^{t-1}} + (1-\frac{n-1}{E^{t-1}})\frac{q}{N-1}\right) + P({k_i}^{t-1} = n)\left(1-\frac{n}{E^{t-1}} - (1-\frac{n}{E^{t-1}})\frac{q}{N-1}\right) \\ &\quad + P({k_i}^{t-1} = n+1)\left(p_0\frac{n+1}{E^{t-1}} + q\frac{n+1}{E^{t-1}}\right) \end{aligned} \tag{1}$$

Since the total number of mers in a genome is usually very large, and each mer species only accounts for a very small fraction of the genome, we assume that the copy number of each mer species evolves independently. Therefore, the above equation can be viewed as an approximation of the copy number distribution of all possible mers in a genome. This assumption is validated by Monte Carlo simulations.

We fit our model to the mer frequency distribution in real genomes by numerical simulations. The initial condition is set as a random sequence of length 1kb. The iteration proceeds until the graph size reaches the corresponding size of the real genome under study. The model has only two free parameters, but it is able to fit the distributions of mers over a wide range of scales (Figure 2). Analyses of the model reveals that deletion process is as essential as substitutions and duplications. When deletion is omitted, the model can still fit the 6-mer frequency distribution quite well — a result consistent with Lee, *et al.* [7]. However, this model can no longer fit the frequency distribution of mers of other sizes (Figure 2).

Our model fits not only the distributions of nucleotide words (mers) in genomic sequences but also the distributions of amino acid words (aa) in protein sequences [14]. The results on the amino acid level further proves the essential role of deletion in the model. Therefore, although deletion can be neglected when modeling the distributions of large functional units, such as gene families [6], it

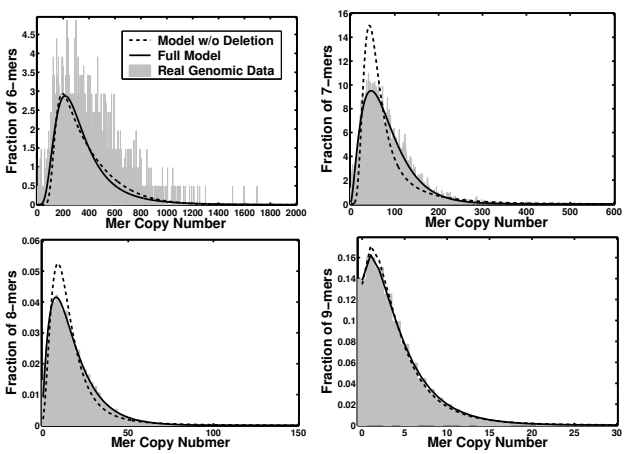

FIG. 2: Our model (black solid lines) is fitted to the distributions of different mer sizes (6, 7, 8 and 9-mers) in *E. coli* K12 genome (gray bars). The results are compared between the full model (black solid lines) and the model without deletion (black dotted lines). Both models fit quite well to 6-mer distribution. However, for other mer sizes (7, 8 and 9), the full model, which includes deletion, obviously does much better than the other, which only incorporate substitution and duplication.

Prokaryotic Genomes	**6-mer**		**7-mer**		**8-mer**	
	q	p_1/p_0	q	p_1/p_0	q	p_1/p_0
M. genitalium	0.0117	1.03	0.0477	1.14	0.1725	1.67
M. pneumoniae	0.0410	1.10	0.1241	1.61	0.3106	2.81
H. influenzae	0.0136	1.03	0.0366	1.10	0.1546	1.58
S. subtilis	0.0095	1.02	0.0320	1.09	0.1406	1.50
E. coli K12	0.0101	1.01	0.0210	1.02	0.0708	1.05
P. abyssi	0.0155	1.02	0.0674	1.09	0.2028	1.51
P. furiosus	0.0111	1.02	0.0303	1.03	0.1132	1.21
S. solfataricus	0.0072	1.02	0.0338	1.05	0.0670	1.15
S. tokodaii	0.0059	1.02	0.0190	1.05	0.0637	1.22
Eukaryotic Genomes	**8-mer**		**9-mer**		**10-mer**	
	q	p_1/p_0	q	p_1/p_0	q	p_1/p_0
S. cerevisiae	0.0568	1.18	0.1944	2.00	0.3704	2.97
C. elegans	0.0112	1.06	0.0350	1.30	0.1307	2.97
A. thaliana	0.0051	1.02	0.0131	1.05	0.0519	1.24
D. melanogaster	0.0114	1.04	0.0393	1.20	0.1728	2.78

TABLE I: Graph model parameters (q, p_1/p_0) fitted to the mer-frequency distribution data (6 to 8-mer for prokaryotic genomes and 8 to 10-mer for eukaryotic genomes) from the whole genome analysis. Different mer lengths are shown for prokaryotes and eukaryotes because of the large difference in their genome sizes.

has a significant effect in modeling the statistical features at a smaller scale. The diminished role of deletions on gene family level may be due to the strong selection pressure against deletions of large sizes. But in a scalable and more generalizable model, deletion remains irreplaceable.

It is worth noting that the model parameter q (substitution probability) is significantly lower when fitted to amino acid distributions than to the distributions of mers of corresponding sizes (three times the amino acid size). Such differences can be explained by the purifying selection in coding regions and the degeneracy of amino acid codons. The successful application of the model on amino acid frequency distributions imply an expected, yet important phenomenon — the evolution processes and their resulting statistical structures on the genomic level are well-reflected on protein level. Naturally, we expect our model to generalize, in order to explain the statistical features in higher-level genomic or cellular processes, such as protein-protein interaction networks, signaling pathways, etc.

In our empirical studies, the model is applied to various mer lengths and to genomes from organisms of various domains: eubacteria, archaea, unicellular and multicellular eukaryota. The fitted values of the two free parameters (q and $\frac{p_1}{p_0}$) in some of the studied genomes are listed in table I.

The fitted parameter values in the table show some interesting properties. First, the substitution probabilities (q) increase monotonically with the mer length (l) in each genome. This may reflect the scaling effect introduced by fixing the size of duplications and deletions in the model as the size of one mer (l). However, in the related biological processes, while one substitution always changes one mer to another, the size of a duplication or deletion event may be larger than the mer size in the model, leading to changes in the copy numbers of more than one mer. For a duplication or deletion event of a certain size, when the mer size increases, the number of mers effected by the event decreases. Therefore, the relative probability of substitution of longer mers tend to be bigger than those of shorter ones. Second, the model fits various distributions nicely when p_1/p_0 is set to be larger than 1, and the values of p_1/p_0 grow with the mer lengths in each genome. These results validate our assumption ($p_1 > p_0$), but also suggest that the probability of duplication decays more slowly than the probability of deletion when the length of the duplicated/deleted fragment increases. Therefore, duplication events of large sizes are more likely to happen than deletion events of large sizes. Since the model parameters scale with the corresponding mer lengths, it is possible to deduce the distributions of the actual sizes of duplication and deletion events in a particular genome [1] when the model is fitted for a sufficiently large number of mer-sizes. Third, although not always, the relative substitution rate q as well as the ratio p_1/p_0 tend to be anti-correlated with the genome size (in the table, genomes in different domains are listed according to their genome sizes in an ascending manner). These observations have a natural explanation if one expects the sizes of the fragments in both duplication and deletion events to be bigger in larger genomes.

The fitted model parameters to a genome provide estimators for the relative frequencies of substitution, du-

plication and deletion events over the evolution history of the genome. For example, *A. thaliana* genome has been reported to [10] have gone through several rounds of large-scale duplication events, accompanied by massive gene loss relatively recently. In contrast, no recent large-scale duplications or deletions are detected in *S. cerevisiae* [15], *C. elegans*, or *D. melanogaster*. Consistent with those genome studies, in *A. thaliana* the relative substitution rate q and the ratio between duplication and deletion p_1/p_0 are much lower compared to other eukaryotic genomes, indicating duplication and deletion events of higher rate and larger scale.

The genomes in the table are separated into prokaryotics and eukaryotics. The prokaryotic genomes are further divided into eubacteria (upper half) and archaea (lower half). It is interesting to notice that although the parameter values vary among the organisms from the same domain, the most dramatic variations are observed between prokaryotic and eukaryotic genomes. More specifically, in eukaryotic genomes, one observes a discernible reduction in the relative substitution rates (q) as well as in the ratios between duplication and deletion probability (p_1/p_0) for a certain mer length. We conjecture that this might be explained by how efficiently the basic evolutionary mechanisms operate at the molecular level and how they differ between prokaryotes and eukaryotes, or between haploid and diploid genomes; these basic evolutionary mechanisms, in turn, are determined by many processes such as DNA repair efficiency, recombination rate, and tolerance of deletions or insertions.

In spite of the important role that natural selection plays on evolution of genomes, our model is still capable of explaining distributions at various scales and in different organisms without implicitly modeling selection force. This may imply that most of the events during genome evolution are actually neutral. A more interesting implication is that natural selection acts not only on individual gene level, it may also act by tuning the relative frequencies of the basic stochastic processes (deletion, duplication and substitution) in evolution. In that case, it is likely that the variation in the model parameter values across different organisms further reflects the differences in the organisms' interaction with their environment. For example, *M. genitalium* and *M. pneumoniae* are both parasitic microbes, but one lives in primate genital system, and the other in respiratory tracts. Although they belong to the same genus, the estimated duplication frequency vs. deletion frequency ($\frac{p_1}{p_0}$) is much higher in *M. pneumoniae* than in *M. genitalium*. Consistent with this observation, genome comparison study [11] has revealed that *M. pneumoniae* genome contains an ortholog to every gene in *M. genitalium*, but it also has extra copies of genes for cell envelope and DNA restriction.

The general and unifying nature of our model suggests a universal minimal set of mechanisms (deletion, duplication and substitution) that are driving genome evolution. Ultimately, these basic schemes can be viewed as the results of selection not just on genomes, but also on the processes modulating their evolution. These processes persist possibly because their combination balances the plasticity against the robustness of not just the genomes, but also the cellular and inter-cellular structures. These features of genomic processes hold the answer to how genomes can be both stable, and yet paradoxically mutable and adaptive.

[1] Y. Zhou and B. Mishra, *Lecture notes in computer science* (2003), in Press.

[2] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein, Genome Biology **3**, RESEARCH0040 (2002).

[3] H. Jeong, B. Tombor, R. Albert, Z. N. Oltavai, and A. L. Barabasi, Nature **407**, 651 (2000).

[4] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Olitvai, Nature **411**, 41 (2001).

[5] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).

[6] I. Yanai, C. J. Camacho, and C. DeLisi, Physical Review Letters **85**, 2641 (2000).

[7] L. Hsieh, L. Luo, F. Ji, and H. C. Lee, Physical Review Letters **90**, 018101 (2003).

[8] D. L. Hartl, Nature Reviews Genetics **1**, 145 (2000).

[9] N. L. Johnson and S. Kotz, *Urn Models and Their Application* (John Wiley & Sons, 1977).

[10] T. J. Vision, D. G. Brown, and S. D. Tanksley, Science **290**, 2114 (2000).

[11] R. Himmelreich, H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann, Nucleic Acids Research **25**, 701 (1997).

[12] To avoid the complication of inversions, we treated two inversely complimentary mers as one species. (For example, 5'-ATCG-3' and 5'-CGAT-3' are counted as one mer species, i.e., their frequencies are combined.) Therefore, for mer size l, there are $\frac{4^l}{2}$ species of l-mers.

[13] We approximate the probability of a specific mer being chosen to substitute another mer during substitution as $\frac{1}{N-1}$ (instead of $\frac{1}{3l}$). This approximation stands when mer size l is small.

[14] When the model is applied to amino acid (aa) frequency distributions, each node in the graph represent a different peptide of length w, and there are 20^w nodes in the graph. The initial condition is a random peptide sequence of 300 single amino acids

[15] Although S. cerevisiae also went through a large-scale duplication, the event is far more ancient, and the duplicated segments have significantly diverged.