Appl Biochem Biotechnol (2010) 162:321–328
DOI 10.1007/s12010-009-8793-7

A Law of Mutation: Power Decay of Small Insertions and Small Deletions Associated with Human Diseases

Jia Zhang · Li Xiao · Yufang Yin · Pierre Sirois · Hanlin Gao · Kai Li

Received: 26 May 2009 / Accepted: 24 September 2009 /
Published online: 10 October 2009

Abstract Indels in evolutionary studies are rapidly decayed obeying a power law. The present study analyzed the length distribution of small insertions and deletions associated with human diseases and confirmed that the decay pattern of these small mutations is similar to that of indels when the mutation datasets are large enough. The describable decay pattern of somatic mutations may have application in the evaluation of varied penetrance of different mutations and in association study of gene mutation with carcinogenesis.

Keywords Indels · Cancer · Length distribution

Introduction

Mutations are the substrates of evolution as well as of genetic diseases and cancers [1–5]. Single nucleotide substitution is the leading type of genetic variation [6–8]. As compared to single nucleotide substitution, insertion mutation, deletion mutation, and

J. Zhang · L. Xiao
Clinical Molecular Diagnostic Center, the Second Affiliated Hospital of Soochow University, Suzhou 215004, China

Y. Yin
SNP Institute, University of South China, Henyang 421001, China

P. Sirois
Department of Pharmacology, University of Sherbrooke, Sherbrooke, QC J1H5N4, Canada

H. Gao
Beckman Research Institute, City of Hope, Duarte, CA 91010, USA

K. Li (✉)
Department of Pharmacology, Medical College, Soochow University, Suzhou 215123, China
e-mail: kaili34@yahoo.com

some mutations colocalize, the inserted and deleted sequences are less frequent but may have more pathological implications particularly when they are in the coding regions [9–11]. In evolutionary studies, insertion, deletion, and mutations containing both inserted and deleted sequences are generally termed as "indel." In germline and somatic studies, however, indel usually means the type of mutations colocalize both inserted and deleted sequences [12–14].

For indels between aligning orthologous sequences, a power law or modified power law was described, with the power parameter varying from 1.5 to 2.3 [15–18]. This length distribution, together with the estimated mutation rate, provides new algorithm in the determination of the role of natural selection in molecular evolution and in many bioinformatic applications such as phylogenetic tree reconstruction and gene discovery [19–26]. Interestingly, Zhang and Gerstein reported that indels in human pseudogenes and introns rapidly decays by the power law [27] similar to the length distribution pattern in indels between orthologous sequences comparison between species, which prompted us to examine whether there is a describable decay pattern of small insertions and deletions associated with human diseases [28, 29].

To address (1) whether there is a describable decay pattern of insertions and deletions linked with human diseases, and (2) whether the length distribution of inheritable mutations and somatic mutations is different, the present study performed two independent assays using relatively large mutation sets were chosen in. The first assay analyzed the length distribution of all the small insertion and small deletion collections deposited in the Human Genome Mutation Database (HGMD). The second assay analyzed the length distribution of the small insertion and deletions in a hypermutable gene of TP53.

Materials and Methods

Databases Retrieved

The HGMD (www.hgmd.org, July 2006) and IARC TP53 Mutation Database (www.p53.iarc.fr, February 2007) (IARC, International Agency for Research on Cancer) were chosen based on their large depository items. The size distribution of the small insertions and deletions in the HGMD was directly obtained from the database. The number of small mutations in the IARC TP53 mutation database was calculated using Microsoft Excel.

Approximation of the Power Coefficients of the Fitting Curves

The trends of the distributions were fit with all available ways by Excel in order to approach the best fitting curves. Three types of regressions were performed whenever applicable: (1) single regression over the original data up the size of 20 nucleotides; (2) consecutive regression from the size of one to six nucleotides to sizes of one to 20 of the genetic variants; and (3) in-frame mutation excluded regression. In the case of consecutive regressions, their exponents were further analyzed with a linear regression.

Results and Discussion

The present study analyzed two of the largest human mutation databases and revealed that small insertion and deletion mutations rapidly decay with the square of number of nucleotides

inserted or deleted. The decay pattern can be expressed as $M_{(i/d)N} = M_{(i/d)1} \times N^{-2}$. Here, the M represents the number of mutations, i/d represents insertion or deletion, respectively, and 1 or N represents the length of nucleotide(s) inserted/deleted. Thus, small insertions/deletions within a limited range of sizes ($M_{(i/d)N}$) can be well estimated using the events of one base insertion or deletion ($M_{(i/d)1}$) that are the most frequent among insertion and deletion mutations. The events of small insertions/deletions are negatively related to the size of the mutated fragments in power decay with exponent of 2.

Length Distribution of the Small Insertions and Deletions Deposited in HGMD

Regression analysis showed that a power decay fits better than all other possible decay patterns for the small insertions and small deletions covering the size from one to 20 nucleotides directly retrieved from the HGMD among the decay patterns tested. A similarity in the power coefficients of the decay curves of small insertions/deletions is found between the small insertions and small deletions in the germline mutation database from a collection of more than 1,000 genes (Fig. 1). The incidence of deletions are more frequent than insertions in these small pathological mutations, which is consistent with what observed in revolutionary studies from neutral indels, and may suggest that there are some common quantitative features between the neutral indels in revolutionary studies and the pathological mutations analyzed in the present study.

Two interesting observations were obtained by comparing the observed and the simulated incidences of the indels. First, the most significant difference is between the three nucleotide deletions and insertions. The simulated and the observed incidences of the three nucleotide deletions were about the same; whereas the three nucleotide insertions observed were significantly less than the estimated (106 observed vs 246 estimated, $p<0.0001$). The second interesting observation is the generally higher incidences of the observed indels longer than ten as compared to the simulated, which imply that the decay curves fit better for smaller indels. Based on the second observation, a consecutive regression was then performed for small insertions/deletions from one to four to one to 20 stepwise. The exponents from these consecutive regressions were further analyzed by linear regression (Fig. 1b), which confirmed the hypothesis that the shorter indels fit the decay curves better than the longer ones.

Although the net effect of naturally occurred indels is to reduce genome size, the effect on genome size from insertion and deletion mutation is different for the noncoding region and coding region. In noncoding regions, the deletions are about three to four times more than insertions, while the insertions are even more than the deletions in the coding region in mammalian genome [30]. Whether there is a relationship between the less harmful effect of adding one amino acid to a protein and the protein expansion during evolution is to be elucidated. It is possible that some proteins expanded through the genetic code tinkering by adding one amino acid gradually. A randomness testing of indels might be used in the evaluation of positive or negative selection during evolution among genetic alignment among species.

Length Distribution of the Small Insertions and Deletions Deposited in IARC TP53 Mutation Database

When the insertions/deletions deposited in the IARC TP53 database were plotted, simple regression of the observed density of small insertions/deletions against their size confirmed that the number of small mutations rapidly decayed with the increase of the

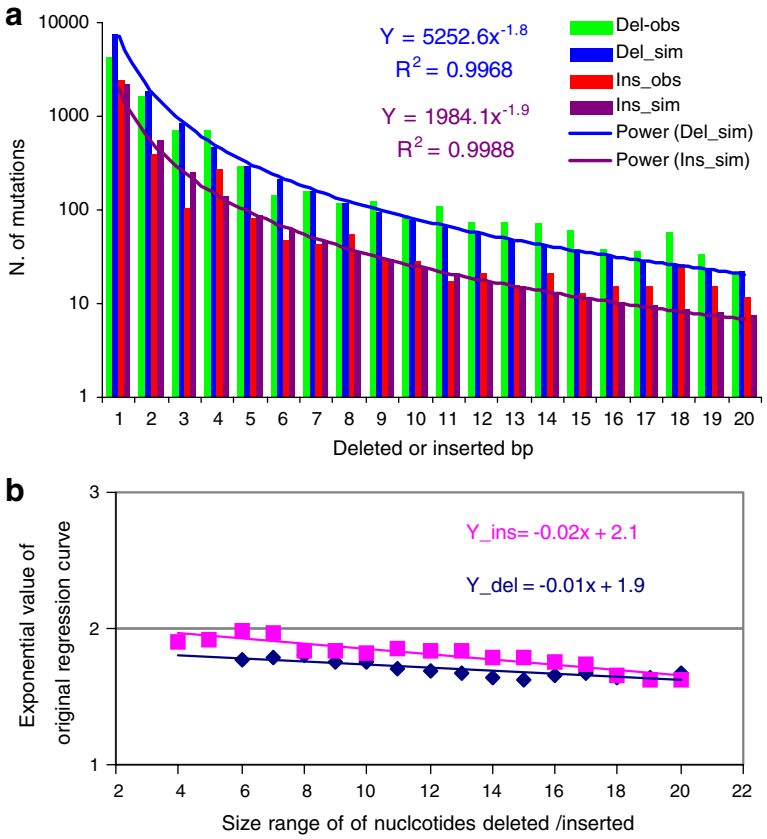

Fig. 1 Size distribution of small insertions and deletions deposited in HGMD. **a** Size distribution of the small mutations observed and those simulated according to the fitting formula with an exponent near 2 for both insertions and deletions. When the size distribution curves were compared with simulated decay curves, a prominent discrepancy was identified at the insert/delete three nucleotides. Whereas inserted three nucleotides matches closely to the simulated, the number of three nucleotide deletions observed is significantly less than the simulated, indicating that the addition of an amino acid to a protein is less deleterious as compared to deleting an amino acid from a protein. **b** The double regression of the primary exponents of the decay patterns. The exponents approach *2* as the mutation size decreases which suggests that the decay pattern is more accurate for small mutations

sizes of the mutations (Fig. 2). As the decay patterns varied upon to the size of the mutations, a double-regression algorithm by consecutive regression of their size distributions of these mutations allowed the approximation of the value of the exponents of being 2. The double-regression assay revealed some additional interesting results after differentiating mutations with and without in-frame insertions/deletions. Both the small insertions and deletions decay faster for frameshift mutations as compared to those of in-frame mutations. It is possible that frameshift mutations and the in-frame mutations possess different carcinogenic ability. The output from double-regression, exponents of being 2, explicates the limitation of length distribution pattern from the quantitative analysis of somatic mutations: the power decay may be different for in-frame mutations and frameshift mutations; and may be different for relatively shorter mutations and relatively longer mutations.

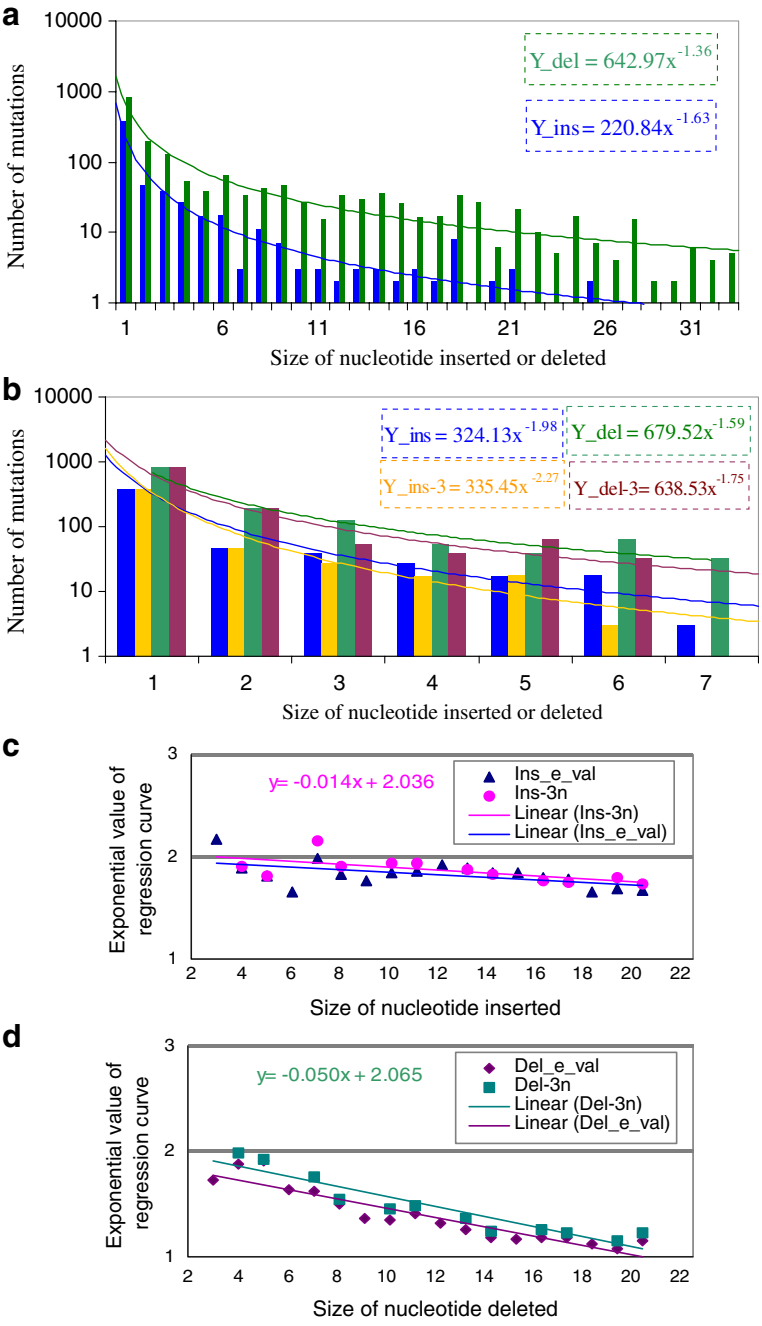

Fig. 2 Size distribution of small insertions and deletions of human TP53 somatic mutation. **a**, **b** Direct plotting of the density of the small mutations against the nucleotide inserted or deleted at the sizes up to *31* nucleotides and up to *7* nucleotides, respectively. The exponents were close to 2 for both the insertions and deletions when the regressing region shortened from *31* to *7* nucleotides. This is even more significant when in-frame mutations were excluded. **c**, **d** Linear regression of the primary exponents from the consecutive power regressions for the observed small insertions and deletions

The TP53 gene is hypermutable, with mutations reported nearly in every triplet code but the penetrance of the p53 mutant is smaller than one as evidenced by the different ages of developing cancer in members of the3 Li–Fraumeni family [31]. At the molecular mechanism level, a single mutation in one allele of the TP53 gene is not sufficient enough to initiate the carcinogenesis. A TP53 mutation is neither crucial nor sufficient for carcinogenesis which is in agreement with the facts that about 50% of human cancers are p53+/+ and that the p53−/− knockout mice do survive [32, 33]. The randomness pattern of TP53 mutations may be attributed to their penetrance of less than one and the hypermutable property of the gene itself. A gene possessing mutations with a random pattern in cancer tissue is to be confirmed by linkage analysis.

Although TP53 and some other inherited gene abnormalities such as the APC, BRCA1, and BRAC2 are associated with the increased lifetime risk in developing a cancer, most of cancers are associated with the accumulation of somatic mutations. The Catalogue of Somatic Mutations in Cancer database based on large-scale resequencing of human genes identified an average of ten to 100 somatic mutations in each individual tumor. In breast and colorectal cancers, Sjöblom et al. identified 189 genes (average of 11 per tumor) that were mutated at significant frequency [34]. Whether a gene harboring somatic mutations is associated with the carcinogenesis or development of a cancer in specific tissue is difficult to address and usually with no clear answer about the genotype–phenotype relationship as compared to inherited genetic abnormalities including cancers elucidated by means of linkage analysis such as the case of Li–Fraumeni syndrome. The present study suggest a determination of randomness of mutagenesis of individual gene according to the decay pattern of small mutations: If the mutations of a specific gene identified in a specific cancer tissue occur nonrandomly, a direct association between the gene and the particular cancer could be suggested. However, some restrictions may apply to the randomness testing: (1) the mutations are in sufficient numbers, and (2) the bases inserted or deleted are relatively short. One point must be kept in mind is that if the mutations of a specific gene identified in a specific cancer tissue occur randomly, randomness testing cannot be applied to reject its association with carcinogenesis.

Acknowledgments The authors are indebt to Dr. Gabriel Gutiérrez at Departamento de Genética, Universidad de Sevilla, Sevilla, Spain for proofreading this manuscript. This paper is partially supported by Department of Personnel Jiangsu province “Liu Da Ren Cai Gao Feng” grant (07-B-033), and The National Natural Science Foundation of China (No. 30970877).

References

1. Kamb, A. (2003). Mutation load, functional overlap, and synthetic lethality in the evolution and treatment of cancer. *Journal of Theoretical Biology, 223*, 205–213.
2. Sommer, S. S. (1994). Does cancer kill the individual and save the species? *Human Mutation, 3*, 166–169.
3. Temin, H. M. (1988). Evolution of cancer genes as a mutation-driven process. *Cancer Research, 48*, 1697–1701.
4. Hughes, A. L. (2008). Near neutrality: leading edge of the neutral theory of molecular evolution. *Annals of the New York Academy of Sciences, 1133*, 162–179.
5. Pfeifer, G. P., & Besaratinia, A. (2009). Mutational spectra of human cancer. *Human Genetics*, *24*. (Epub ahead of print)
6. Ott, J., & Hoh, J. (2001). Statistical multilocus methods for disequilibrium analysis in complex traits. *Human Mutation, 17*, 285–288.

7. White, P. S., Kwok, P. Y., Oefner, P., & Brookes, A. J. (2001). 3rd international meeting on single nucleotide polymorphism and complex genome analysis: SNPs: 'some notable progress'. *European Journal of Human Genetics, 9*, 316–318.
8. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature, 409*, 928–933.
9. Ball, E. V., Stenson, P. D., Abeysinghe, S. S., Krawczak, M., Cooper, D. N., & Chuzhanova, N. A. (2005). Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human Mutation, 26*, 205–213.
10. Chen, J. M., Chuzhanova, N., Stenson, P. D., Férec, C., & Cooper, D. N. (2005). Complex gene rearrangements caused by serial replication slippage. *Human Mutation, 26*, 125–134.
11. Chuzhanova, N. A., Anassis, E. J., Ball, E. V., Krawczak, M., & Cooper, D. N. (2003). Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human Mutation, 21*, 28–44.
12. Scaringe, W. A., Li, K., Gu, D., Gonzalez, K. D., Chen, Z., Hill, K. A., et al. (2008). Somatic microindels in human cancer: the insertions are highly error-prone and derive from nearby but not adjacent sense and antisense templates. *Human Molecular Genetics, 17*, 2910–2918.
13. Gonzalez, K. D., Hill, K. A., Li, K., Scaringe, W. A., Wang, J. C., Gu, D., et al. (2007). Somatic microindels: analysis in mouse soma and comparison with the human germline. *Human Mutation, 28*, 69–80.
14. Gu, D., Scaringe, W. A., Li, K., Saldivar, J. S., Hill, K. A., Chen, Z., et al. (2007). Database of somatic mutations in EGFR with analyses revealing indel hotspots but no smoking-associated signature. *Human Mutation, 28*, 760–770.
15. Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., & Hein, J. (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Research, 18*, 298–309.
16. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature, 428*, 493–521.
17. Chang, M. S., & Benner, S. A. (2004). Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *Journal of Molecular Biology, 341*, 617–631.
18. Gu, X., & Li, W. H. (1995). The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *Journal of Molecular Evolution, 40*, 464–473.
19. Cartwright, R. A. (2006). Logarithmic gap costs decrease alignment accuracy. *BMC Bioinformatics, 7*, 527.
20. Kim, J., & Sinha, S. (2007). Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics, 23*, 289–297.
21. Lunter, G. (2007). Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics, 23*, i289–i296.
22. Yamane, K., Yano, K., & Kawahara, T. (2006). Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. *DNA Research, 13*, 197–204.
23. Denver, D. R., Morris, K., Lynch, M., & Thomas, W. K. (2004). High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature, 430*, 679–682.
24. Lunter, G., Ponting, C. P., & Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology, 2*, e5.
25. Halpern, A. L., & Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution, 15*, 910–917.
26. Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution, 26*, 473–480.
27. Zhang, Z., & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research, 31*, 5338–5348.
28. Li, K. (2006). Small insertions and deletions is revealed in association with the number of inserted or deleted nucleotides. *J Nanhua University, 34*(1–2), 9.
29. Li, K., Xiao, L., Yin, Y. F., & Zhang, J. (Oct 9-13, 2006) How to associate the somatic mutations and a specific cancer. *56th ASHG*, New Orleans, USA.
30. Taylor, M. S., Ponting, C. P., & Copley, R. R. (2004). Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Research, 14*, 555–566.

31. Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Jr., Nelson, C. E., Kim, D. H., et al. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science, 250*, 1233–1238.
32. Donehower, L. A. (1996). The p53-deficient mouse: a model for basic and applied cancer studies. *Seminars in Cancer Biology, 7*, 269–278.
33. Hollstein, M., Sidransky, D., Vogelstein, B., & Harris, C. C. (1991). p53 mutations in human cancers. *Science, 253*, 49–53.
34. Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science, 314*, 268–274.