J Mol Evol (2005) 60:274–275
DOI: 10.1007/s00239-004-0109-5

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

On Slippage-Like Mutation Dynamics Within Genes: A Study of Pseudogenes and 3′UTRs

Manami Nishizawa, Kazuhisa Nishizawa

Department of Biochemistry, Teikyo University School of Medicine, Kaga, Itabashi, Tokyo 173-8905, Japan

Received: 6 April 2004 / Accepted: 1 October 2004 [*Reviewing Editor:* Dr. Dmitri Petrov]

Recent analyses show that even very short repetitive sequences are prone to slippage-like mutations. Characterization of the dynamics of such mutations should increase our knowledge about the background frequencies of extension and contractions commonly occurring within genes, allowing the effect of the selections on particular repetitive motifs to be assessed. Consideration of the slippage-like changes may also help the reconstruction of the phylogenetic tree of a gene. In our previous report (Nishizawa and Nishizawa 2002), we described a method which finds the best alignment between two sequences by performing simulations of various types of changes including slippage-like ones. The method can be used to optimize the "slippage-controlling parameters" so that they explain the observed evolution of sequences well. While we believe that the presented framework was reasonable, we acknowledge that the accuracy of the parameter estimation in the paper was limited due to a dearth of informative pseudogenes at the time. The purpose of this letter is to present (1) the result of the pseudogene analysis that supplements our previous report and, in addition, (2) the results of 3′UTR analyses, which have implications for evolutionary changes of 3′UTR sequences.

We extended the analyses to 160 human and 120 mouse pseudogenes, which were randomly chosen from the recent database (Echols et al. 2002; Zhang et al. 2004). The total numbers of insertions and deletions ("the indel-total") that we estimated to have occurred for human and mouse pseudogene/functional-gene pairs were 627 and 931, respectively. Based on the simple binominal model, these numbers can be considered to produce only the negligible sampling errors, and therefore the *p*-values in the following primarily reflect the potential error in our likelihood estimations.

Our method simulates the changes of the sequences in a manner such that the various patterns of insertions and deletions can be examined (Nishizawa and Nishizawa 2002). We have categorized the slippage-like changes into the three types of changes that are accounted for by the parameters ζ_1, ζ_2, and ζ_3, respectively. The ζ_1 parameter indicates the probability, for the insertion events scheduled during the simulation, that each insertion becomes a duplication (instead of an insertion of random nucleotides), as represented by the example ATCAGC → ATCA<u>CA</u>GC or ATCA<u>GC</u>GC, where the introduced 2-nt segment is underlined. (If, for example, ζ_1 is set to 0.6, the simulation is performed such that 60% of insertions [of any lengths] are duplicative, where 30% are the duplications of the upstream nucleotides and the remaining 30% are those of the downstream nucleotides.) Our previous results suggested that human pseudogenes are more likely to undergo the ζ_1-type slippages than those of murids. In fact, further analyses show no difference between human and murids. For both the human and the mouse pseudogenes, the likelihood profile for ζ_1 shows peak at $\zeta_1 = 0.5$, or∼50% are duplicative, indicating that pseudogenes of human and mouse do not have an appreciable difference in ζ_1 -mutations.

The ζ_2 and ζ_3 parameters specify the "extra" probability (compared with nonrepetitive sequences) that the very short repetitive sequences, such as GGGG and CACACA, are subjected to the extension and contraction of the repeat, respectively. (For

Correspondence to: Manami Nishizawa; *email:* nmanami@aol.com

example, the vertical lines in G|G|G|G are called the "modulo-*1* rep-centers," whereas those in CA|C|A|-CA are called the "modulo-*2* rep-centers." When ζ_2 is set to, say, 10, the simulation is modified such that the modulo-*i* rep-centers, where *i* is the length in nucleotides to be inserted upon the insertion event, are assigned the 11 (=1 + 10)-fold probability to "receive" the insertion event compared to the sites without repetitiveness; the increment (=10) is supposed to be the contribution solely by duplicative insertions. ζ_3 similarly deals with the effect of the repetitiveness on the positional preferences of the deletion events. (Nishizawa and Nishizawa 2002). We found that for ζ_2 and ζ_3, both human and mouse pseudogenes show a peak at 7.5, suggesting that there is no significant difference in the probability of slippage-like changes of repetitive segments compared to nonrepetitive ones.

To examine the possible involvement of slippage-like mutations in evolutionary changes of 3′UTR sequences, we extended the analyses to 84 orthologous pairs for human/monkey (*Macaca, mulatta*) and 176 for mouse/rat (*Rattus norvegicus*). For human–monkey and rat–mouse comparisons appropriate reference groups are not available, therefore, we assumed the parsimony rule as described by Hein (1989) and inferred the ancestor sequences using our method. Hence, the polarity between insertions and deletions was not determined, and instead, the indel-total was considered. The results are summarized as follows. (i) Pseudogenes and 3′UTRs were largely similar in terms of the proportion (among the indel-total) of them that is accounted for by each of the ζ_1-, ζ_2-, and ζ_3-type slippages. (ii) For the primate and murid 3′UTRs, the indel-total (normalized per 1000 substitution events) was 196.4 and 144.5, respectively, indicating that murids tend to undergo a greater number of total indel events per a fixed number of substitutions than primates ($p < 0.0001$). (iii) The frequencies (per 1000 substitutions) of the *short* (1-nt) indel vs the *long* (5- to 20-nt) indel were 78 vs 13.5 for the human 3′UTRs, while the corresponding values for the pseudogenes were 6 vs 2.6 (insertions) and 19 vs 10 (deletions). This indicates that the frequency of the indels of *short* segments as compared to the indels of *long* ones is higher for the 3′UTRs than for the pseudogenes ($p < 0.001$). This is most likely due to the constraints on the length of the 3′UTRs. (Let us note that the pseudogenes have distinct ages, so it is unreasonable to compare the absolute number of events between the pseuodogenes and the 3′UTRs.)

Regarding result (ii), we also found that the lengths of the 3′UTRs appear to change more frequently than those of the pseudogenes on a per-substitution basis ($p < 0.0001$), possibly because of positive selection. (The indel-total in the human pseudogenes was only 57.0 per 1000 substitutions.) This explanation is supported by result (iii), which implies that the constraint on the length *per se* is stronger for 3′UTR than for pseudogenes. Results (i) is also interesting in that those changes which the slippage-like changes create in 3′UTRs may largely provide sufficient fitness to the 3′UTRs; it could be that the very specific changes of sequence are not necessary and the length adjustment is sufficient.

By introducing more specific parameters, our method showed, for example, that GGG/CCC is more susceptible to contractions and extensions than AAA/TTT (our unpublished result). Such a fine parameter tuning may eventually lead to accurate characterization of the dynamics of very short repetitive sequences. In light of the usefulness of the gaps within aligned sequences for finding the correct phylogenetic tree (McGuire et al. 2001), it is our hope that such parameters will be utilized, for example, to choose one tree topology from the candidate topologies that are regarded as equally likely based on conventional approaches.

References

Echols N, Harrison P, Balasubramanian S, Lscombe NM, Bertone P, Zhang Z, Gerstein M (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 30:2515–2523

Hein J (1989) Anew method that simultaneously aligns and reconstruct ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol Bio Evol 6:649–668

McGuire G, Denham MC, Balding DJ (2001) Models of sequence evolution for DNA sequences containing gaps. Mol Biol Evol 18:481–490

Nishizawa M, Nishizawa K (2002) A DNA sequence evolution analysis generalized by simulation and the Markov chain Monte Carlo method implicates strand slippage in a majority of insertions and deletions. J Mol Evol 55:706–717

Zhang Z, Carriero N, Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20:62–67