NOTE: This article in the present form had been submitted to editorial office of ***Genomics*** on August 11, 2009 (GENO-D-09-00174). It sat there for 16 full weeks. On December 1, 2009 I sent a query to the editorial office asking whether there was anything I could do to help the review process. In response, within an hour, the associate editor sent me the complete rejection of the article based on only a single very superficial review. Therefore, I have decided to publish it on my website on December 2, 2009. **I should like to ask the reader to consider August 11, 2009 as the submission date. As a further documentation of the date of publication, the manuscript was deposited on Google Doc on Dec.30, 2009.**

The evolution of Alu mutations in the human genome.

Guenter Albrecht-Buehler
Department of Cell and Molecular Biology
Feinberg School of Medicine
Northwestern University
Chicago, IL 60611
Tel.: 312-503-4261
E-mail: g-buehler@northwestern.edu

Keywords: Alu-elements, base substitutions, retro-transposons, evolutionary age, dynamics, age determination

Running title: evolution of Alu mutations

SUMMARY

The article reports the frequency distributions of the mutants of Alu elements in the human genome. They are remarkably complex and, yet, almost identical for each chromosome, suggesting a universal mechanism for base substitutions in Alu-elements and, possibly, other retro-transposons as well. Conceivably, these mutations of the Alu-elements effectively reduced or even crippled their proliferation which,

otherwise, might have fragmented and destroyed the host genome. The article proposes a simple mathematical model that simulates the observed distributions and offers

(a) a quantitative reconstruction of the evolutionary past of the number of Alu-mutants,

(b) a determination of the times in their past when new copies of the Alu-elements inserted with full capacity to proliferate ('seedings'), arguably giving rise to new sub-families,

(c) a new and simple determination of the evolutionary age of Alu-mutants and, thus, of a minimal age of the domains of the host chromosome in which they were found.

INTRODUCTON

Alu elements and other retro-transposons should pose a lethal threat for every genome they invade. Their method of amplification via transcripts that reinsert into the host genome through reverse transcription [1, 2, 3] could conceivably lead to an exponential 'explosion' of copy numbers that would completely fragment and thus destroy the host genome. In the case of the Alu retro-transposon this catastrophe did not happen to our ancestral genomes it invaded, although its copy numbers in the e.g. human genome exceed 1 million [4, 5]. Lucky for us, our ancestral genomes appear to have found an effective defense strategy that limited the proliferation of the Alu-elements to harmless levels and, in the process, may even have created a selective advantage [6].

One of the defense strategies may have been the mutation of the Alu-elements, possibly aimed at crippling their ability to proliferate. Since the entire spectrum of conceivable point mutations is consistent with the interpretation that all point mutations were caused by auto-mutagenic mechanisms [7], it would make sense, if the genomes had unleashed this arsenal for their defense. As will be shown in this article, the Alu mutants in the human genome of 50 or even more base substitutions outnumber the 'original' Alu-copies by a wide margin. Considering that the Alu-elements are only approximately 280 bases long, such large numbers of base substitution must have had a substantial impact on their functionality.

One might expect that random Alu-proliferation in the host genome followed by random base substitutions of each Alu sequence results in poorly reproducible, rather chaotic distributions of Alu-mutants. Surprisingly, however, the process created precisely defined frequency distributions of Alu-mutants that were the same for all human chromosomes (and chimpanzee chr.1) and depended only on the specific family to which the Alu-element belonged. In order to explain this finding, the present article offers a simple mathematical model of the dynamics of the proliferation of Alu-elements while their capacity to proliferate is

increasingly inhibited by point mutations. If correct, this model will permit to reconstruct the evolutionary past of the Alu mutants and also to predict their evolutionary future. It may even serve to justify the interpretion of Alu mutants as time stamps on the host genome.

The present article adopted several simplifying methods and strategies in order to depict Alu-elements in an easily recognizable way, and to minimize computation times.

1. Characterization of Alu-mutant by the number of their base substitutions regardless of their position.

The most significant simplification used here was the focus on the number of mutations in an Alu-sequence regardless of their position. In view of the high level of sophistication of today's sequence analysis of Alu-elements [4, 8] this approach may seem rather crude. However, similar to aerial photography, the omission of details may sometimes offer a depiction of large-scale features that might otherwise go undetected. I hope to convince the reader that this loss of sequence details, nevertheless, helps describing the overall dynamics of Alu-evolution in its past and future.

2. The restriction to the main Alu families

As a further simplification the article focuses only on the 3 major Alu-families, AluY, AluS and AluJ, while ignoring their division into a total of 217 sub-families [8, 9]. Nevertheless, my search program for Alu-mutants treated the members of all sub-families as mutants and, therefore, did not omit them.

RESULTS

1. The distribution of Alu-mutants in the human genome

A. The genome pixel image (GPxI) of Alu-elements and their mutations

Instead of describing the mutated Alu sequences as strings of letters, we will present them as optical images by the GPxI method described earlier [11]. Briefly, the method assigns to the bases of a DNA sequence the following gray-tone values: A: black, G: white, C: dark gray and T: light gray. This assignment is, of course, arbitrary, but must remain the same throughout. It transforms the consecutive bases of the sequence into a continuous line of pixels with varying gray values. In addition, the method requires the choice of an arbitrary, but also fixed image width W. Whenever the line of pixels reaches W, it wraps around like any other text would, and continues at the beginning of the next line immediately underneath.

It is, of course, also possible to choose the image width W equal to the size of the depicted sequences. In this way, the sequences (e.g. the Alu-sequences) can be written in register, as was done in all GPxIs shown in this article.

The resulting images permit intuitively clear decisions whether a sequence is or is not a mutated Alu-element, while avoiding rather abstract mathematical homology computations. For example, the GPxIs of the sequences of the 217 Alu subfamilies [8] if placed in register

appear as shown in Fig. 1a. Obviously, no detailed sequence analysis is required to recognize in the GPxI that all these sequences all are variations of essentially the same motif.

In order to find the Alu-mutants in the human genome, the article uses the search program 'GA-dnaorg.exe' written by the author, and specific search primers of 209 [b] size whose GPxIs are shown in Fig. 1b. Their sequences are listed in Materials and Methods.

B. The choice of the search parameters.

The search algorithm used in the search program was a simple base-by-base comparison between a search primer and a genome sequence while the primer was moved along the genome. The success of the search was defined as a match where the number of base substitutions remained below a certain threshold N. Whenever the program found a suitable match it recorded its position, sequence and exact number of base substitutions in a data file.

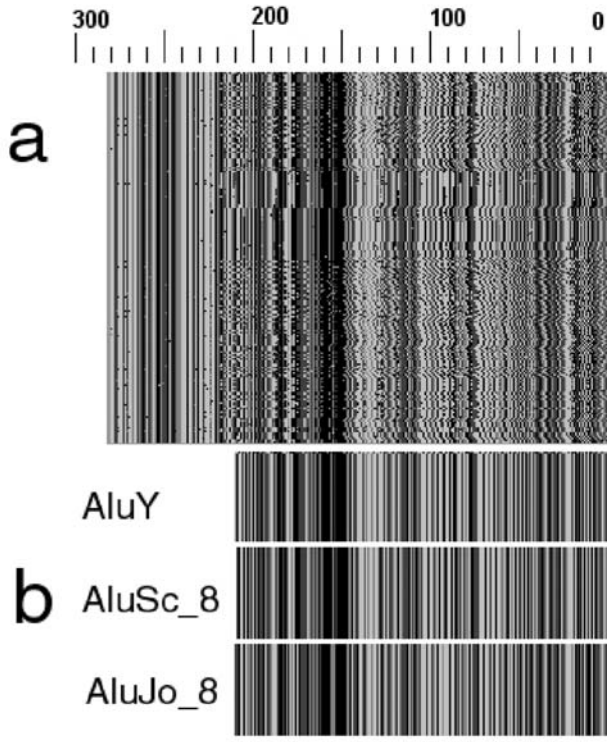

Figure 1.

Genome pixel images (GPxI) of the Alu-sequences used in this article (for the method of GPxI see [11]). The scale on top represents the positions of the bases beginning with the down-stream end of the Alu-sequences.

(a) GPxI of the sequences representing the 217 different Alu subfamilies as listed in [8).

(b) Representative GPxI of 50 identical sequences representing the AluY, AluS, and AluJ families used in this article. They were used as search primers (size = 209 bases) in the search program for Alu-mutants.

The threshold N must not be too small, lest the search would miss too many Alu-mutants. It must also not be too large, lest the search would accept sequences that could no

longer be considered Alu mutants. As shown by their GPxIs (Fig 2) the patterns of the sequences identified by the search program were, indeed, easily recognizable Alu-mutants even for values as high as N = 100 base substitutions, provided the size search primer was 200 bases or larger.

The same criterion of yielding recognizable Alu-patterns was applied to the selection of a suitable size π of the search primers. While a search primer with π = 200 bases searched with a threshold of N = 100 yielded clear Alu-patterns (Fig. 3b), a search primer with π = 50 did not yield recognizable Alu-patterns, even if the threshold N was a low as 25 bases (Fig. 3c).

These and similar criteria led to the choices of a search primer size π = 209 or 213 bases and a threshold of an acceptable number of N = 100 bases substitutions throughout the following.

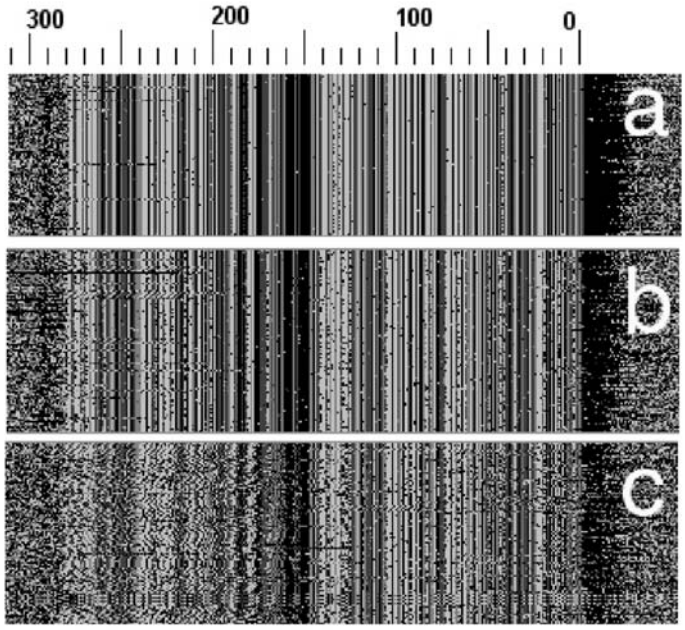

Figure 2.

Effect of the maximal number of tolerated base substitutions on the Alu-mutant sequences found by the search program. The scale on top represents the positions of the bases of each found sequence beginning with the down-stream end of the AluY-sequences in the human genome. The GPxIs show small portions of the upstream and downstream flanks of the various Alu-mutants. Note the appearance of poly-A stretches (=black pixel stretches) at the start of the each down-stream flank of each Alu-sequence found by the search program. The images display the GPxI of 100 AluY-sequences obtained by the search program using the AluY search primer (size 200) and tolerating

(a) up to 5 base substitution.

(b) up to 25 base substitutions.

(c) up to 100 base substitutions.

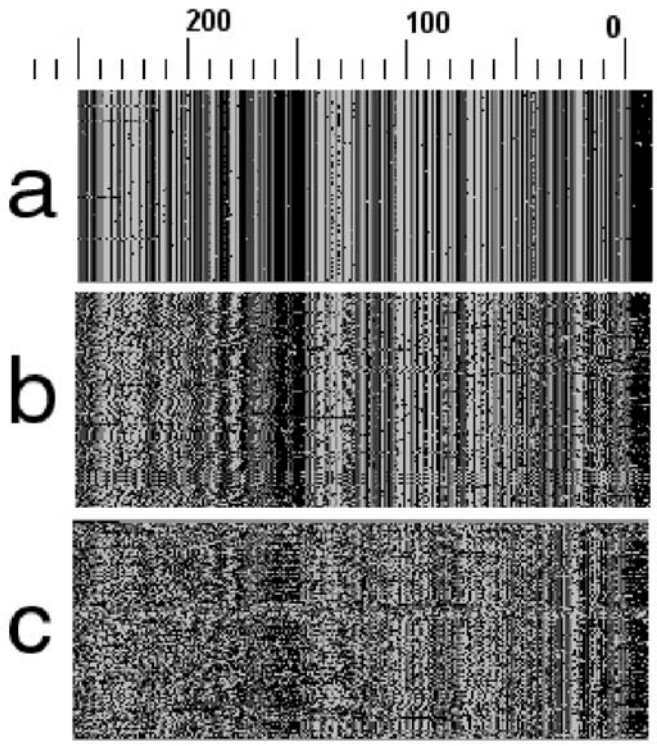

Figure 3.

Effect of the size of the search primer on the Alu-mutant sequences found by the search program. The scale on top represents the positions of the bases of each found sequence beginning with the down-stream end of the AluY-sequences in the human genome. Note that the search program will find and display the entire Alu-sequence, even if the search primer is a shorter sequence.

The images display the GPxI of 100 AluY sequences detected with a search primer size of

(a) 200 bases and up to 5 base substitution.

(b) 200 bases and up to 100 base substitutions.

(c) 50 bases and up to 25 base substitutions.

C. The universal frequency distribution of Alu-mutants.

Applying the described search method individually to the human chromosomes 1 – 22 and X yielded 389,956 AluY mutants. If normalized for the same chromosome size, their frequency distributions were remarkably identical for all chromosomes (Fig. 4a) as evidenced by the very small standard deviations between the values of different chromosomes (bars in Fig. 4).

Similarly, the search program found 171,066 AluS mutants and 172,240 AluJ mutants in human chromosomes 1 – 7. Although their average distribution curves were characteristically different for the different members of the Alu family, different chromosomes yielded again surprisingly identical distribution curves (Fig. 4b, c). The distribution curve of the AluJ mutants consisted almost exclusively of heavily mutated elements, confirming that the AluJ–elements are the oldest of the family [8, 9].

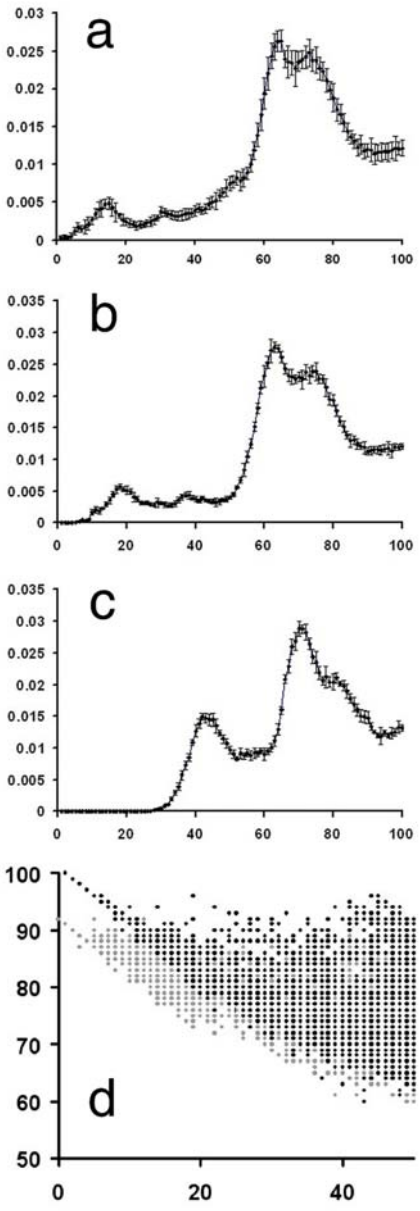

Figure 4.

The remarkably high degree of reproducibility of the mutant distributions of Alu elements in the human genome. Bars indicate the standard deviations of the values of the tested number of chromosomes. Abscissa: number n of base substitutions of the Alu-mutants; ordinate: Frequency of the various Alu-mutants with exactly n base substitutions, normalized to a maximum amplitude of 100.

(a) Distribution of AluY-mutants (averaged over chromosomes 1 – 22, and X)

(b) Distribution of AluS-mutants (averaged over chromosomes 1 – 7)

(c) Distribution of AluJ-mutants (averaged over chromosomes 1 – 7)

(d) Homologies (Needleman-Wunsch algorithm [12]) between all AluY-mutants tested against the AluY search primer (black dots) and the AluJ search primer (gray dots)

Testing the homologies of all recorded AluY mutants against the AluY- and AluJ search primers by the Needleman-Wunsch algorithm [12] yielded characteristically different values for mutants with less than 30 base substitutions (Fig. 4d). However, in case of 40 or more base substitutions both search primers yielded the same cloud of values between 50% and 95% homology. This result is not surprising as one should expect that the distinction between the more heavily mutated elements of the 3 Alu-families AluY, AluS and AluJ may become blurred, as 40-50 or more of base substitutions are likely to wipe out many of the relatively small sequence differences between the 2 search primers (see Figure 1).

A surprising feature was the appearance of multiple peaks in the distributions, suggesting that there had been several waves of increased replication in the evolutionary past of the Alu-elements.

In all cases, the frequency distributions showed a pronounced dominance of Alu-mutations with 50 and more base substitutions over Alu-elements that contained fewer than 50 mutations. Equating large numbers of base substitutions with large evolutionary age, it suggests that most Alu-elements in the human genome are quite 'old'.

Testing chimpanzee chromosome 1 yielded the same distributions as the human chromosomes.

D. The decision, which of 2 Alu-elements is more similar to the 'original' based on their mutant distribution.

The search for mutated Alu-elements poses a fundamental question. How can we know whether the sequence AluY that we used as a search primer is the 'original' sequence? Why should not another mutant Alu-sequence $AluY_m$ with (say) m base substitutions be the 'original' while AluY was one of ***its*** m-fold mutants?

To be sure, there is clear evidence that Alu sequences are part of the 7SL RNA gene of numerous species [10]. However, among them only certain primates have processed it into a retro-transposon, whereas Xenopus and drosophila have highly analogous 7SL RNA genes but no Alu-elements. Therefore, there was early in the evolution of these primates a mutation of the primate 7SL RNA gene or an invasion from the 7SL RNA gene of another species that laid the foundation of the Alu-elements as we know them today.

Obviously, we can never decide whether a particular Alu-sequence is the 'true original' which may no longer exist today. Therefore, in the literature many authors placed the terms 'original' or 'source' sequences in inverted commas as was done in the present article.

Nevertheless, based on the set {M} of all Alu-mutants known today, it is quite possible to determine which of 2 Alu-mutants is more similar to the 'original' than the other. Traditionally, the students of Alu-elements have solved the problem by detailed studies of homologies between domains of different Alu-sequences, which can determine which

sequence pre-dates the other and, thus identify the earliest among them as the most 'original' [4, 8].

The mutant distributions presented here offer another rather simple way to tell which of two Alu-mutants is more similar to the 'original' sequence. Consider the set {M} of mutants that all arose from a common original sequence Alu_0 in the human or any other genome. Using Alu_0 as a search primer will yield a specific mutant distribution $A_0[n]$ similar to the ones in Fig. 4.

Now select one of the mutant Alu_0-sequences X which, unbeknownst to you, differs from Alu_0 by m base substitutions. Using X as a search primer will yield **its** mutant distribution $A_X[n]$ from the same set {M} of Alu-mutants.

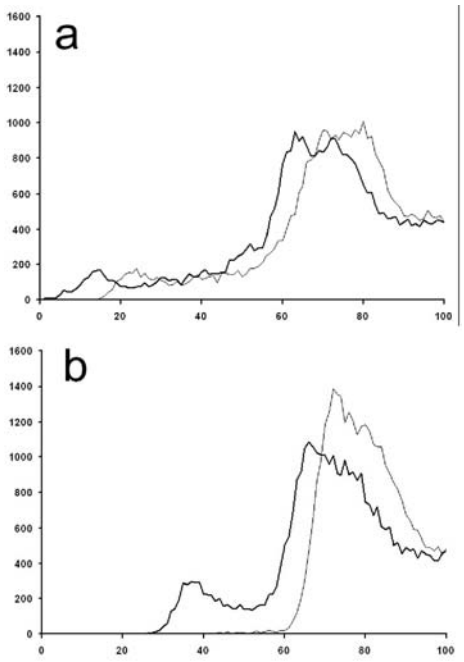

Figure 5.

Dependence of the mutant distribution on the number of initial base substitutions contained in its search primer. All distributions were established from human chromosome 1. The search program using the different search primers allowed up to 100 base substitutions for each mutant. Abscissa: number n of base substitutions of the Alu-mutants; ordinate: absolute count of the various Alu-mutants with exactly n base substitutions.

(a) Mutant distributions resulting from the original AluY search primer with 0 (dark line) and 15 (gray line) additional base substitutions.

(b) Mutant distributions resulting from the original AluY search primer with 30 (dark line) and 63 (gray line) additional base substitutions.

The more mutations the search primer has suffered, the more its mutant distribution curve looses mutants with low numbers of base substitutions and shifts to the right.

The comparison between the distributions $A_0[n]$ and $A_X[n]$ will show quite easily that the search primer Alu_0 is more similar to the 'original' Alu-sequence than X by the following criteria: *Compared to $A_0[n]$ the distribution $A_X[n]$ will be shifted towards large numbers of base substitutions while lacking mutants of X with 1, 2, 3,… and other small numbers of base substitutions.*

For example, Fig. 5 shows the mutant distributions obtained from certain AluY-mutants X(i) with i = 0, 15, 30, and 63 base substitutions which were used as search primers on human chr.1 Clearly, the more base substitutions the search primer X(i) contained, the fewer mutants could be found that contained less than (say) 30 base substitutions.

The finding can be explained by the fact that {M} contains many mutants of Alu_0 with 1, 2, 3,… and other small numbers of base substitutions, as the presence of such mutants is the definition of {M}. Hence, $A_0[n]$ will contain substantial numbers of mutants with fewer than 30 base substitutions.

On the other hand, it is extremely unlikely to find in {M} a single sequence that could qualify as a mutant of (say) X(30) that contains additional 1, 2, 3,… or other small numbers of base substitutions. Such a sequences would have to be mutants of Alu_0 with the exact same 30 base substitutions as X(30) in exactly the same positions, but contain 1, 2, 3,… additional ones in other (or the same) positions. The probability to find such mutants is very small since the number of all possible mutants with 30 base substitutions is on the order of 10^{60}, while {M} contains only a miniscule fraction of them, namely approximately 10^6. Hence, $A_{X(30)}[n]$ will contain almost no mutants with only 1, 2, 3, … base substitutions.

This reasoning was used earlier, in order to conclude that AluJ was much older than AluY because it contained almost no mutants with fewer than 30 base substitutions (Fig. 4c). Likewise, one can see immediately that the age of AluS is in between AluY and AluJ but more similar to AluY, as its mutant distribution has fewer such mutants than AluY but many more than AluJ.

2. Mathematical model of the dynamics of Alu mutations

The described invariance of the frequency distributions and the predominance of heavily mutated Alu-elements are quite counter-intuitive results and, therefore, need explanation. The following mathematical model of Alu-mutation will try to provide one.

The model introduces only the most obvious variables and parameters that are needed to discuss quantitatively the evolutionary fate of the Alu-mutants. Furthermore, the equations of the model are kept simple and comply with common sense. For all these reasons I included them in the main body of the text, instead of banning them into an Appendix.

A. The basic variables and equations.

Definitions:

L: size of the 'original' Alu-element

R: number of recursions of computation. It plays the role if 'time' in the model and will be calibrated in terms of evolutionary time T (see equ. 6)

ΔR: number recursions between successive computations (usually, ΔR=1)

T: calibrated evolutionary time

ΔT: calibrated time intervals of computation corresponding to 1 recursion of computation (approx. 250,000 yrs; see equ.6)

P: number of recursions needed to develop the model into the present

n: number of base substitutions in an Alu-mutant

N: maximal number of base substitutions allowed ($n \leq N$; $N = 100$) by the search program

A[n,R]: number of Alu-mutants that contain exactly 'n' base substitutions at 'time' R

Using these definitions the model considers only the effects of base substitutions on a particular 'original' Alu-element. It calculates at successive recursions R the number of mutants A[n,R] that contain exactly n base substitutions at 'time' R. The article will describe R interchangeably as 'evolutionary time' or as 'recursions'. The values of R will start with R = 0 (i.e. the first appearance of the Alu-element in the genome) and proceed to R = P (i.e. the present time). The consecutive time points are assumed to be spaced by equal distances of 'ΔR '. As mentioned earlier the numbers of mutation n must remain below the threshold 'N', i.e. the maximal number of mutations allowed which leave the mutated Alu-element still recognizable as a mutant of the original sequence.

During the time interval ΔR the number of mutants A[n,R] will change for 3 reasons.

(a) Some of the mutants replicated (called 'replication[n,R]'),

(b) Some of the A[n-1,R] mutants received a base substitution and added to the number A[n,R] (called 'gain[n,R]'), and

(c) Some of the A[n,R] mutants received a base substitution and moved up to the next category A[n+1,R] (called 'loss[n,R]').

Hence, we can write as change of A[n,R]

{1} $$\Delta A[n,R] = \{ replication[n,R] + gain[n1,R] - loss[n,R] \} \cdot \Delta R; \qquad (0 \leq n \leq N ;\ 0 \leq R \leq P)$$

The simplest assumption about the replication[n,R] is that it is proportional to the number of mutants A[n,R] and to a probability z[n] that expresses the capacity of an Alu-mutant with n base substitutions to replicate:

{2} $$replication[n,R] = \alpha \cdot z[n] \cdot A[n,R] ;$$

Similarly, we assume that gain[n,R] and loss[n,R] are proportional to the numbers of mutants A[n-1,R] and A[n,R] respectively from which they originated. Furthermore, they should be also proportional to certain probabilities v[n] and w[n] that describe how likely an Alu-mutant with n-1 or n base substitutions will receive an additional one. Hence,

{3a} $gain[n,R] = \beta \cdot v[n] \cdot A[n-1,R];$

{3b} $loss[n,R] = \beta \cdot w[n] \cdot A[n,R];$

with the constant β describing the probability of a base substitution

As to the replication probability z[n], its detailed properties do not matter as long as it vanishes very rapidly with the number n of mutations. Otherwise, every solution of the above equations would be equivalent to an explosive growth of the copy numbers of the original Alu-element and all its mutants. Hence, I chose the simplest inactivation function

{4a} $z[n] = \exp(-n/\gamma)$; *with* γ *the inactivation constant*

As to the gain and loss probabilities, it stands to reason, that the probability of a mutant to receive one more base substitution is proportional to the fraction of the not yet mutated bases. Hence,

{4b} $v[n] = (L-n-1)/L;$

{4c} $w[n] = (L-n)/L;$

Beginning with R = 0 and n = 0 one needs to compute the values of ΔA [n,0] for each value of n and subsequently obtain the values of A [n,R] by the recursion

{5} $A[n,R+\Delta R] = A[n,R] + \Delta A[n,R]; \qquad (0 \le n \le N;\ 0 \le R \le P)$

Equ.3 ignores the possibility of more than one simultaneous base substitution. Otherwise, if 2 or 3 of them would occur during the same time interval ΔR, then one would have to include the contributions from A[n-2,R] and A[n-3,R] to the gain of A[n,R].

However, this scenario is extremely unlikely. It is known that Alu-elements appeared around 60 million years ago [8, 9]. During this time the majority of them accumulated on average some 60 base substitutions (see Fig. 4), which suggests an approximate frequency of 10^{-6} base substitution per Alu per year. Therefore, the occurrence of 2 or 3 simultaneous substitutions would have probabilities of 10^{-12} and 10^{-18} per Alu and year, and may be neglected.

Technical note: The mathematically trained reader will have noticed that the above equations 1 – 4 describe a set of linear difference equations for which there are well established, elegant methods of solution. Unfortunately, it is difficult to convert their parameters such as the Eigenvalues of the coefficient matrix into quantities that have readily understandable biological meaning such as mutation rates, replication rates, etc. Furthermore, as pointed out in the next section, it will be necessary to render this matrix explicitly time dependent. As a result, there will be no explicit solutions of these equations.

Therefore, the present article preferred a more pedestrian way of solving the equations with a computer program called ‘Alu_dnaorg.exe’, which was written by the author and carries out the required recursions of equ.5 one step at a time.

B. The fitting of the observed distributions of Alu mutants.

The mentioned computer program yielded essentially only 2 kinds of non-trivial, realistic solutions. One, which will be called the 'single seeding solution' describes the time course of the mutants A[n,R] following a single 'infection episode with a number of 'original' or 'source' sequences (called a 'seeding'). The other describes the time course of the mutants A[n,R] following multiple episodes of new seedings at different times, and will be called 'multiple seeding solution'.

a. Single seeding solution

Assuming a single seeding of 'original' Alu elements I used the simulation program to determine the simplest, most fundamental solution of the equ. 1 – 5. I selected as parameters a size of the Alu-elements of L = 200, an initial fraction of 'original' Alu-elements of A[0,0] = 1.8 [%], an inactivation constant of γ = 5 [mutations], and values for the (dimensionless) mutation and replication terms of $\beta \cdot \Delta R$ = 0.51, and $\alpha \cdot \Delta R$ = 0.435. The solutions are shown in Fig. 6 as a function of the number of recursions. Their actual calibration in units of evolutionary time will be deferred to Section 3.

It appeared that the basic single seeding solution is represented by a single narrow peak, reminiscent of a Gaussian. In contrast to a Gaussian, however, it is not symmetrical about its maximum. As time increases, the amplitude of the peak grows while it migrates to the right, i.e. towards larger numbers of base substitutions. This behavior of the single seeding solution may not be immediately obvious. However, it may be made plausible as follows.

The steady growth in numbers of Alu mutants follows because replication can only increase the numbers of Alu-mutants. Once these numbers are large enough, even a very low probability of replication can still increase them, albeit by relatively small increments. Nevertheless, these increments continue to add up to larger and larger numbers.

As to the 'migration to the right', it follows because the number of base substitutions can only increase if one disregards the unlikely possibility that a base substitution accidentally reverts a mutation back to its original base. Therefore, the number of Alu-sequences that still contain only a small number of base substitutions decreases steadily, while the ones with the largest number of such mutations acquire even more.

b. Multiple seeding solution

In contrast to the single seeding solution, the actual frequency distributions of the Alu-mutants in the human genome showed several peaks (Fig. 4), suggesting that several seeding episodes of 'fresh' (i.e. fully replicative) Alu-elements had occurred in the past [13, 14, 15, 16]. The simulation program provided for the occurrence of several episodes of new Alu-elements appearing and replicating.

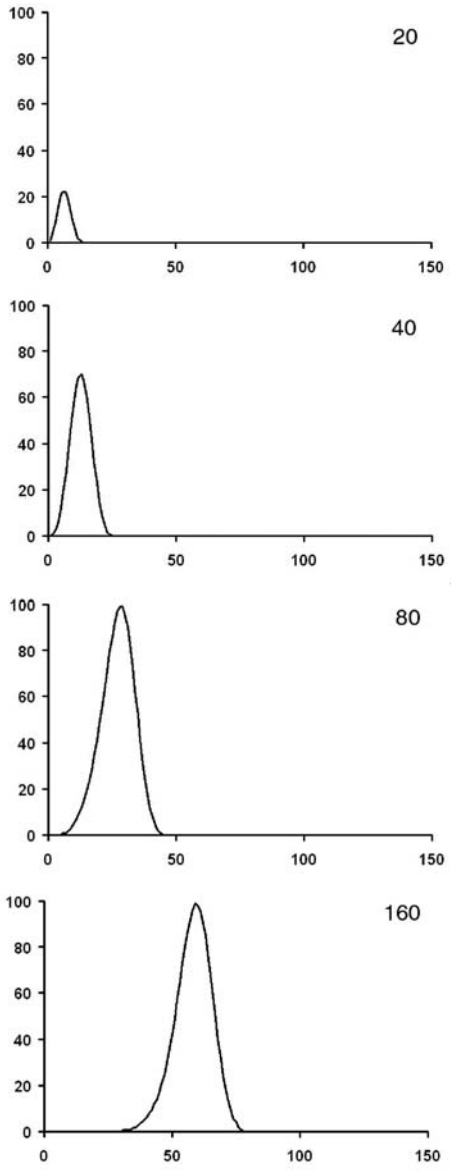

Figure 6.

Illustration of the time development of the simplest mutant distribution that results from the mathematical model ('single seeding solution'). (The number of recursions noted in the right upper corner of each panel). In this case the number of copies of Alu-elements develops in time as a result of replication and mutation after a single seeding of a small number of initial, 'original' elements. Abscissa: number of base substitutions of the mutants; ordinate: frequencies of mutants normalized to yield a maximum of 100 at recursion = 160.

As shown in Figure 7 the mutant distributions of the AluY, AluS, and AluJ mutations could, indeed, be simulated with reasonable accuracy by introducing 5 and 4 individual single seeding episodes at specific recursion number R_i. In section 3 they will be converted into actual times T_i. All parameters used for the simulation are shown in Table 1. It appeared that

each new wave of Alu-elements required its own set of parameters in order to fit the experimental distributions.

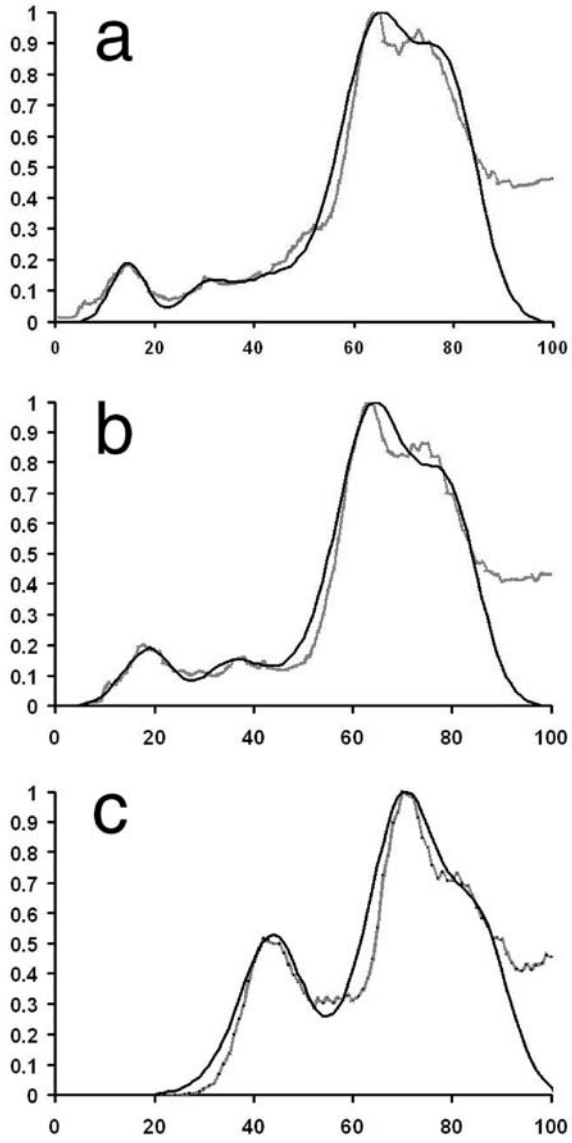

Figure 7.

Fitting of the mutant distributions of AluY, AluS, and AluJ in the human genome using the mathematical model ('multiple seeding solution') The gray lines show the experimental distributions (cf. Fig. 4). The fitted curves are drawn in dark lines. The fitting required the introduction of 5 and 4 different seedings. The parameters used for the fitting are listed in Table 1.

The reader will have noticed that the experimental distributions differed considerably from the simulated ones at levels of 90 or more base substitutions. The reason for this discrepancy is not a failure of the model but an inevitable consequence of the search method.

Above certain high levels of mutations the search program will increasingly misidentify unrelated sequences as heavily mutated Alu-sequences. This becomes obvious in the extreme case, where the search program would accept all of the 280 bases in the Alu-element as mutations. In this case, *every* contiguous sequence of 280 bases contained in the chromosome would qualify as an Alu-mutant and would be counted. In other words, every experimental distribution inevitably rises to the very high level of (chromosome size)/280 counts as it approaches abscissa values of 280 mutations.

c. The need for initial mutations of newly seeded Alu mutants.

There is no reason to assume that the 'fresh' Alu-elements of the various seedings were necessarily identical to the original sequence. It is conceivable that they contained already one or several base substitution. Indeed, the success of the fitting procedure required several of the most recent seedings to have one or several initial mutations. They are listed in Table 1.

3. The calibration of the evolutionary age of Alu-elements

A. Conversion between recursions and evolutionary time.

As mentioned above, the number of recursions R was used in the mathematical model of Alu-mutation in place of a time variable. In order to calibrate it in terms of the actual evolutionary time T the article made the same assumption as the field in general [8,17], namely that the increasing number of base substitutions occurred at a steady rate. (That is NOT to say that the Alu-mutants occurred at a steady rate. On the contrary, their numbers jumped up several times at the times of the different 'seedings'.).

Therefore, the calibration needs only 2 time points and their corresponding numbers of recursion to determine the conversion factor τ between evolutionary time and number of recursions. The obvious choice for the first point was the first appearance of Alu-elements (T_0 = -60 million years [6, 8]), corresponding to $R_0 = 0$, and as the second point the present time ($T_{pres} = 0$). Its corresponding number of recursions R_{pres} was derived from the fitting of the AluJ distribution. Consistent with other determinations, it was obviously the oldest Alu-element, because it contained practically no members with less than 30 mutations (Fig. 4c). To fit it with the mathematical model required 240 recursions (Table 1), yielding $R_{pres} = 240$. Hence, the conversion factor became $\tau = 60/240 = 1/4$ [million years/recursion]. It was the same value for all other Alu-elements, although their value of R_{pres} was different (see below). In other words the time increment $\Delta R = 1$ of the mathematical model in real time spans 250,000 years.

{6} Let T = evolutionary time, R = number of recursions, and R_{pres} = final number of recursions which creates the present day distribution of mutants, then

$$T\ [10^6\ yrs] = \tau\,(R - R_{pres}),\quad with\ \tau = 0.25\ [10^6\ yrs/recursion]\ ;$$

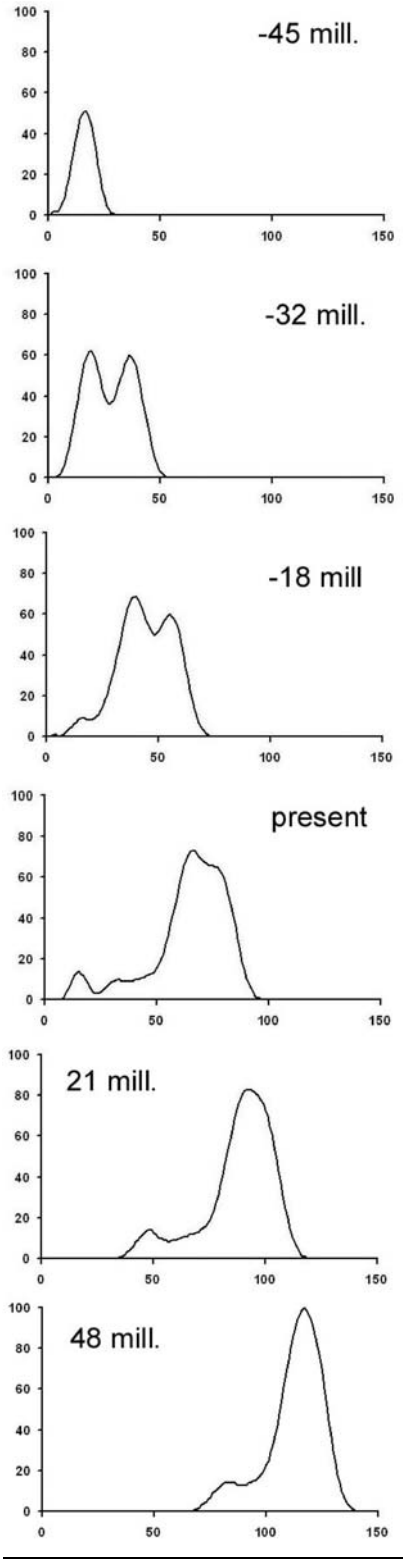

Figure 8.

The time development of the distribution of the AluY-mutants in the human genome, reconstructed and predicted by the mathematical model and calibrated by equ.6.

As an immediate application of equ.6, Figure 8 shows the progression of the simulation of the AluY mutations calibrated in evolutionary time from its beginning into the far future. In this case we used R_{pres} = 200 (see Table 1) for the calibration, because the distribution of the AluY mutants required 200 recursions to reach the present time.

B. The transition from simulated numbers of mutants to actual ones.

Each simulated distribution curve represented a total number of S simulated mutants. Therefore, the actual numbers of Alu-mutants can be calculated by multiplying each value of A[n,T] with the same scale factor λ.

{7} Scale factor $\lambda = M / S$, where

M = the total number of experimentally observed Alu-mutants, and

S = $\sum_n A[n,P]$, the total number of mutants simulated for the present time P.

This approach is legitimate because the solutions of equations 1 are obviously scale invariant i.e. if A[n,T] (0≤n≤N) is a solution then $\lambda \cdot A[n,T]$ is also one for any arbitrary constant number λ.

In the case of the human genome there were 389956 AluY-mutants (see Results 1B) and the fitting program yielded a value of S(AluY) = 5296. Hence, λ(AluY) = 389956/5296 = 73.6. The corresponding values for the 2 other Alu-families were λ(AluS) = 72.7 and λ(AluJ) = 72.3.

The scale factor λ must also be applied to the initial numbers of every seeding. For example, the normalized initial number of A[0,0] = 1.8 for AluY becomes a value of $A[0,0] \cdot \lambda$ = 1.8·73.6 = 133 initial copies of the 'original' AluY. Applying the values for the accuracy of the parameters (Table 1, last column) the mathematical model claims that some 60 million years ago our ancestral genome was 'invaded' by only 133 ± 27 copies of the original AluY element.

C. Seedings of Alu mutants and the appearance of Alu-subfamilies.

Applying equ.6 to the times when new seedings appeared in the mathematical simulation (Table 2, column A) yielded the values shown in Table 2 column B. The appearance of new, major Alu sub-families [8, 9] is listed in column C of the Table. The remarkably close correspondence between the values in columns B and C suggests that the seedings may be interpreted as the appearance of new copies of Alu elements that have the full (original) proliferation capacity and, thus, gave rise to new sub-families. In other words, the mathematical model expresses the appearance of new Alu sub-families as new seedings. Based on this interpretation Fig

Figure 9 shows for the 3 main Alu families AluY, AluS, and AluJ the times and relative numbers for the new seedings and, thus, the appearance of new sub-families.

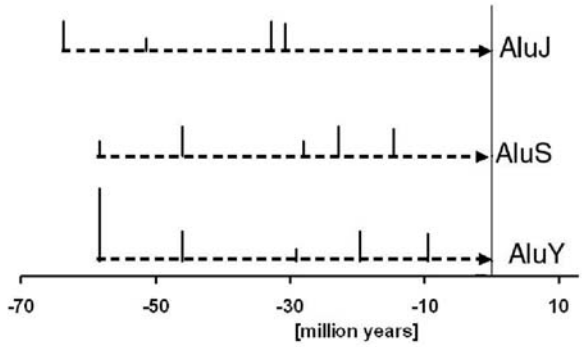

Figure 9.

Timing and relative magnitude of the various seedings of Alu-elements in the evolutionary past of the AluY, AluS, and AluJ families reconstructed by the mathematical model. They coincide rather well with the appearances of new Alu sub-families [8, 9].

DISCUSSION

1. Implications of the mathematical model

A. The basic assumptions of the model.

The mathematical model proposes that the number of Alu-mutants A[n,T] that contain n base substitutions at any given time T

(a) increases through replication, although the probability of replication diminishes rapidly with n,

(b) increases because some of the (n-1)-fold mutants acquire one new base substitution proportional to the fraction of their un-mutated bases,

(c) decreases because some of the (n)-fold mutants acquire one new base substitution proportional to the fraction of their un-mutated bases, and

(d) increases through occasional 'seedings' of Alu-elements with full replicative capacity.

These rather minimal assumptions which comply with simple common sense were sufficient to reproduce the details of the actual mutant distributions of AluY, AluS and AluJ to a high degree of accuracy. None of theses assumptions expresses any chromosome specificity. Therefore, the mathematical model explains the main finding of this article, namely the remarkable similarity between the Alu-mutant distributions of different chromosomes.

B. The quantitative reconstruction of the evolutionary past of Alu-mutations.

A major attraction of the mathematical model is the possibility to reconstruct the past mutation distributions and their future (Figure 8 and 9). Of course, the model cannot predict whether and when future seedings may occur.

The list of the past seedings may not be complete. It seems possible to obtain a more accurate simulation of the mutant distributions than illustrated in Fig. 7 by adding several further 'minor' seedings that introduced only very small numbers of 'fresh' Alu-elements. These additional seedings may explain the large number of 217 Alu-sub families [8]. However, they were omitted here in order to offer a more transparent presentation of the mathematical model.

C. The seemingly large number of required parameters.

The simulation of each observed mutant distribution used as many as 28 parameters. The number may appear large and, thus, may seem to render the model less meaningful.

However, one should keep in mind that the length L = 200 of Alu-search primers and the time T_1=0 for the first seeding are not fitting parameters. Furthermore, a number of six parameters is clearly a necessary minimum for each single seeding solution, because the processes of inactivation (which requires 2 parameters), replication and mutation of the Alu-elements, their initial numbers and time of the seeding are all independent of each other and require separate parameters. Eventually, the number of 28 parameters arises as the minimal number of parameters to describe the 5 and 4 separate seedings.

D. The dynamic parameters of Alu-mutation.

As to the replication of all newly seeded Alu-elements, the common value of the inactivation constant γ = 5 suggests that on average a fraction of 1/e = 37% of all newly seeded Alu-mutants were rendered incapable of replication after only 5 base substitutions. Not even all of the remaining elements were replicating, but only 42 % of them (equ.2) because the different Alu-families had a rather similar replication factor $\alpha \cdot \Delta T$ =0.42. However, the reader should be reminded that even very small probabilities of replication may support an impressive increase in total numbers of Alu-elements, provided there are large enough copy numbers already present.

As to the mutation of all newly seeded Alu-elements of the different Alu-families, their parameters were similar enough to equate them to a common value $\beta \cdot \Delta T$ =0.5 for the purpose of the discussion. It suggests that half of the un-mutated fraction w = (L - n)/L of each n-fold Alu-mutant received a base substitution every ΔT = 250,000 years (see equ.6).

The mathematical model also introduced the possibility that each newly seeded Alu-element may have started with a certain number of initial base substitutions (called initial # mutations in Table 1) compared to the 'original' Alu-element. These numbers which are not subject to a scale factor are surprisingly small. However, they have a substantial impact on the fitted distribution.

2. Possible use as time stamp.

Once calibrated in terms of evolutionary time, one can use the mathematical model to determine the age of a particular Alu-mutant Alu_X. What is more, in this way one would also be

able to determine a minimal age for the host chromosome or even a part of the host chromosome in which this particular mutant was found. After all, it stands to reason, that most of the host genome existed before the Alu-element invaded it.

To this end one would use the sequence of Alu_X as a search primer to establish its mutant distribution throughout the entire chromosome. Subsequently, the age of Alu_X can be determined by assessing how many mutants with small numbers of base substitutions are contained in its mutant distribution, as it was described in section 1D of the Results.

3. Applicability to other retro-transposons.

The present article focused on the Alu retro transposons in the human genome, because they are the most numerous and best studied and, thus, could be used for a more detailed discussion of the mathematical model. However, no part of the mathematical model used any specific property of Alu-elements or the human genome that would not also apply to other retro transposons in other species. Therefore, it should be able to model the specific mutant distributions of other retro transposons in other species as well.

MATERIALS AND METHODS

The sequences of the human genome were obtained from the UCSC site. The analysis program, "GA_dnaorg.exe", and the simulation program, “Alu- dnaorg.exe" were written by G.A.-B. using Visual C++ (Microsoft , Redmond, WA).

The search primers for the Alu families J, S and Y were the following sequences:

AluJ primer :

CCCAGGAGTTCGAGACCAGCCTGGGCAACATAGTGAGACCCCATCTCTACAAAA
ATTTAAAAAATTAGCCAGGCATGGTGGCGCATGCCTGTAGTCCCAGCTACTCGGGAGGC
TGAGGTGGGAGGATCGCTTGAGCCCAGGAGGTCGAGGCTGCAGTGAGCTATGATCATG
CCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGTCTC.

AluS primer:

TCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAA
TACAAAAAATTAGCCGGGCGTGGTGGCACGCGCCTGTAGTCCCAGCTACTCGGGAGGCT
GAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGC
CACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTC.

AluY primer:

AGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATA
CAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTG
AGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGC
CACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCT.

ACKNOWLEDGEMENT

The work was supported by the Robert Laughlin Rea endowed chair held by the author.

TABLES

Table 1

	AluY	AluS	AluJ	*)
L (assumed Alu size)	200	200	200	N/A
total recursions T_{max}	220	220	240	6
first seeding:	yes	yes	yes	N/A
second seeding:	yes	yes	yes	N/A
third seeding:	yes	yes	yes	N/A
fourth seeding:	yes	yes	yes	N/A
fifth seeding:	yes	yes	no	N/A
time of seeding T_1	0	0	0	N/A
time of seeding T_2:	46	46	46	6
time of seeding T_3:	110	114	116	6
time of seeding T_4:	146	134	124	6
time of seeding T_5	184	165	N/A	6
Alu[0,0] $_1$:	1.8	1.8	1.8	20%
Alu[0,0] $_2$:	2.0	2.0	2.0	20%
Alu[0,0] $_3$:	0.8	1.0	0.8	20%
Alu[0,0] $_4$:	2.0	2.0	2.0	20%
Alu[0,0] $_5$:	4.8	1.0	N/A	20%
inactivation γ	5	5	5	1
$\beta_1 \cdot \Delta R$:	0.51	0.51	0.51	0.01
$\beta_2 \cdot \Delta R$:	0.51	0.51	0.51	0.01
$\beta_3 \cdot \Delta R$:	0.50	0.50	0.50	0.01
$\beta_4 \cdot \Delta R$:	0.50	0.50	0.50	0.01
$\beta_5 \cdot \Delta R$:	0.50	0.50	N/A	0.01
$\alpha_1 \cdot \Delta R$:	0.43	0.43	0.42	0.01
$\alpha_2 \cdot \Delta R$:	0.43	0.43	0.43	0.01
$\alpha_3 \cdot \Delta R$:	0.45	0.42	0.40	0.01
$\alpha_4 \cdot \Delta R$:	0.46	0.40	0.40	0.01
$\alpha_5 \cdot \Delta R$:	0.46	0.39	N/A	0.01
init.#mutations	0	0	0	N/A
init.#mutations$_2$	0	0	0	<<1
init.#mutations$_3$	1	1	1	<<1
init.#mutations$_4$	2	1	0	<<1
init.#mutations$_5$	2	0	N/A	<<1

*) Accuracy of the parameters. We define the 'accuracy of a fitting parameter' as the absolute value of a change of this parameter (while all other parameters are kept at their best fitting values) that would yield an unmistakable deterioration of the fit.

Table 2

A B C

0	-60	AluJo	(-60)
45	-48.8	AluSx_3	(-44)
110	-32.5	AluS	(-36)
146	-23.5	AluY	(-24)
184	-14	AluYa5	(-16)
240	0	(present)	(0)

A : recursion number at the time of the seedings.
B: calculated time of seeding [mill.yrs] using equ. 6.
C: Subfamily; in brackets (time of origin [mill.yrs]) according to reference 8.

REFERENCES

1. Mathias SL, Scott AF, Kazazian Jr. HH, Boeke JD, Gabriel A (1991). Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808–1810.
2. Rogers, J (1983) Retroposons defined. *Nature* **301:** 460
3. Dewannieux M, Esnault C, Heidmann, T (2003). LINE-mediated retrotransposition of marked *Alu* sequences. *Nat. Genet.* **35:** 41–48.
4. Han K, Xin J, Wang H,Hedges DJ, Garber RK, Cordaux R, Batzer MA (2005) Under the genomic radar: The stealth model of Alu amplification. Genome Res. 15: 655-664
5. Batzer, M.A. and Deininger, P.L.. (2002). Alu repeats and human genomic diversity. Nat. Rev. Genet. 3: 370-379.
6. Britten RJ (1994) evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc. Natl. Acad.Sci. USA 91:5992-5996.
7. Albrecht-Buehler, G. The spectra of point mutations in vertebrate genomes. (2009) BioEssays 31:98-106 (http://www3.interscience.wiley.com/cgi-bin/fulltext/121641840/PDFSTART)
8. Price AL, Exkin E, Pevzner PA. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res. 14: 2245-2252 (doi: 10.1101/gr.2693004
9. Britten RJ (1994) evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proc. Natl. Acad.Sci. USA 91:6148-6150.
10. Ullu E, Tschudi Chr. (1984) *Alu* sequences are processed 7SL RNA genes. *Nature* 312, 171 - 172 ; doi:10.1038/312171a0
11. Albrecht-Buehler, G. Outline of a Genome Navigation System Based on the Properties of GA-Sequences and Their Flanks. (2009) PLoS ONE 4(3): doi:10.1371/journal.pone.0004701
12. Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (3): 443-53.

13. Shaikh TH, Deininger PL (1996) The role and amplification of the HS Alu subfamily founder gene. J. Mol. Evol. 42: 15–21
14. Matera, A.G., Hellmann, U., and Schmid, C.W.(1990). A transpositionally and transcriptionally competent Alu subfamily. Mol. Cell Biol. 10: 5424-5432.
15. Leeflang, E.P., Liu, W.M., Hashimoto, C., Choudary, P.V. and Schmid, C.W. (1992). Phylogenetic evidence for multiple Alu source genes. J. Mol. Evol. 35: 7-16.
16. Shen, M.R., Batzer, M.A., and Deininger, P.L. (1991) Evolution of the master Alu gene(s). J. Mol. Evol. 33: 311-320.
17. Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393-1398.