Genetic Algorithm-based Model of Evolutionary Dynamics of Class II Transposable Elements

HADI QUESNEVILLE* AND DOMINIQUE ANXOLABÉHÈRE

Laboratoire de Dynamique du Génome et Evolution, Institut Jacques Monod,
2, Place Jussieu 75251 Paris Cedex 05, France

(Received on 8 December 2000, Accepted in revised form on 16 July 2001)

We propose a new conceptual framework to study the dynamics of transposable elements. Based on a genetic algorithm, our model is designed as a self-organizing system. Our results show that transposable elements could emerge from a single endonuclease gene. The DNA repair mechanisms appear to condition the emergence success of class II TEs. Antagonist selective forces acting on transposable elements and their hosts induce by their opposition differences in the sequence evolution of the functional domains and of the copies.

© 2001 Academic Press

Introduction

Transposable elements (TEs) are mobile DNA sequences that can be found repeated and dispersed in a genome. All the copies of a TE belong to what is called a TE family. They can be classified according to the mechanism by which they move from one genomic site to another. Class I transposable elements correspond to elements that use an RNA intermediate in their transposition; they are also called retrotransposons. Class II TEs, known as DNA-transposons or DNA transposable elements, use DNA. Within a class II TE family, transposases provided by some intact copies can mobilize defective elements.

Transposable elements have been found in nearly all genomes where they have been sought. They seem to be ubiquitous, and represent a quantitatively important component of genomes (44.4% of the human genome, International Human Genome Sequencing Consortium, 2000). Since the vast majority of the mutations caused by transposons are deleterious, their presence in the genome does not provide any apparent advantage for the survival of the host organism. In spite of their negative effect, transposons can spread and persist in the genome of a species over long periods of time. Intense speculation on how these elements are maintained and spread by virtue of their transposition processes has produced a great deal of theoretical work on their evolutionary dynamics (Charlesworth & Langley, 1989; Brookfield & Badge, 1997; Badge & Brookfield, 1997; Quesneville & Anxolabéhère, 1997b, 1998). Experimentally, many class II TEs have shown an ability to self-regulate by producing a repressor.

Recently, several cellular functions have been found to be closely related to TEs. For example, telomerase is a reverse transcriptase that is related to that of non-LTR retrotransposons. It is not clear whether retrotransposons gave rise to telomerase or vice versa (Nakamura & Cech, 1998; Pardue et al., 1997). Another spectacular example is the origin of antigen-specific immunity. The V(D)J recombination system shares two
main features with DNA transposons: a recombinase encoded by RAG1 and RAG2, and a mobile DNA flanked by recombinase-binding sites. This led to the hypothesis that the immune system comes from the insertion of a TE in the genome of an ancestor of a jawed vertebrate (Agrawal et al., 1998; Hiom et al., 1998). Numerous other examples can be found in Smit (1999).

Very little is known about the origin of TEs, but it is conceivable that it can be traced back to the hypothetical transition from RNA-based genomes to DNA-based ones (Jurka, 1998). There is no doubt that the genomic DNA we observe today has evolved with the close participation of TEs. Many of them have evolved into parasites, but they have probably all kept their properties of “genome builders”. They appear today as crucial actors of genome evolution.

Class II TE transposases have endonuclease properties: they achieve site-specific DNA cleavages (Beall & Rio, 1997). Thus, class II TEs can be thought of as endonuclease genes that recognize the extremities of their own nucleotide sequence. Consequently, TEs could derive from this kind of genes. Moreover, if the first transposase was an endonuclease, its mobile gene needs to generate copies that control their mobility in order to be maintained as a TE family. In this paper, our main concern will be whether TE family can emerge from gene with basic endonuclease properties. To search for the ancestral TE structure, we looked for a minimal organization common to all class II TEs. By the action of protein–DNA interactions and DNA repair, we expect this structure to develop a control mechanism for its mobility.

Here, we propose a new conceptual framework inspired by recent approaches used to study “complex adaptive systems” (Holland, 1995; Forrest, 1993). Based on a specifically designed genetic algorithm, the emergence of a class II TE family is studied as a self-organizing system.

Our results show that a transposable element could emerge from a single gene with basic endonuclease properties. Upon transpositions, this new element can produce mutated and deleted copies. Some of them could interact with other copies in order to reduce the invading capacity of the element. We observe the spontaneous formation of an organized molecular interaction network leading to a controlled mobility. The DNA double-strand break repair process plays an important role in the invasion dynamics, and it appears to condition the emergence success of class II TEs. The distribution of the deletions in the sequences is not centered on the middle of the copies as could be expected given the repair process, but it is shifted to the “endonuclease domain”. The deletion distribution also affects the rate of evolution of the sequence. An antagonism between two selective forces gives rise to a heterogeneity along the TE sequences but also between the different copy types.

Model

Structure of Class II TEs

Figure 1(a) shows the structure of a classical class II TE. The sequence is flanked on each side by a sequence of about 30–250 nucleotides strictly identical and in reverse orientation, called inverted terminal repeats (ITR). Between the ITR, a gene encodes a transposase with two domains: a binding domain that binds the transposase to a target site on the TE sequence near or in the ITR region, and a catalytic domain that cleaves DNA in the ITR. Thus, a minimal TE must have a transposase with a binding domain that we call Rec (for recognition domain) and a catalytic domain that we call Endo (for endonuclease domain). Moreover, the sequence must also have the two transposase target sites that we call IdL and IdR for identity left and right (because these sites determine which transposases are able to recognize this sequence).

The sequence of each of these regions is encoded with a string of four symbols: -, 0, 1, *. For each target site, IdL or IdR, the probability of transposase binding is given by the matching percentage between the target site sequence and the Rec sequence of a candidate transposase. The matches between our symbols are summarized in Table 1. Note that if a sequence accumulates the “*” symbol, it evolves towards a generalization of the recognition between the transposase target site (IdL or IdR) and the transposase recognition domain (Rec); if “1” accumulate, the Rec domain tends to specialize. The probability of cleavage is encoded by the percentage of “1” in the Endo sequence.
FIG. 1. (a) Schematic representation of a typical class II TE, and how the main functional regions are translated in our model; (b) an example illustrating transposase synthesis, transposase recognition and the cleavage in the transposition model; (c) representation of the endonuclease gene used to initialize the simulation of TE emergence.

**TABLE 1**

*Matches between the TE sequence symbols.*

*A “+” indicates a match while a “.” indicates no match.*

<table>
<thead>
<tr>
<th></th>
<th>*</th>
<th>1</th>
<th>0</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**HOST’S REPRESENTATION**

Each individual host is represented by a set of chromosomes. Chromosomes are constituted by two chromatids. Each chromatid contains a finite number of sites which can contain several TE copies. Crossing over occurs between two homologous chromatids, using 1% of recombination between two adjacent sites. Mutations of the TE copy sequences occur during the replication of the chromosome (independent of transpositions) with a probability $m$.

Populations of individuals are modeled with standard population genetic hypotheses: random mating and constant population sizes, i.e. the parents of each zygote are drawn randomly with replacement, zygotes are produced until the offspring population reaches the size of the parent population, and sterile individuals are eliminated from the population.

**TRANSPOSITION**

TE copies are inserted into the various chromatid sites of their host chromosomes. Each
copy produces a transposase, i.e. the binary string of the element reduced to the domains Rec and Endo. The transposase thus produced is stored in a pool of transposases away from the elements. For each element target site, IdL or IdR, the transposase of the pool with the maximum binding probability is tested for its binding to the element. If it succeeds, the transposase is tested for its cleavage of the element using the probability encoded in the Endo sequence. Finally, the element transposes itself if a binding and a cleavage succeed on both sides of the element. Then, the element inserts itself in a new site chosen at random. If one of these steps failed, the element remains in place. For each successful transposition, there is a probability \( s \) that the host becomes sterile. This sterility may be caused by a chromosomal breakage induced by DNA cleavages during the transposition process. Figure 1(b) illustrates schematically an example of this scenario.

When one element transposes, it leaves a gap at the donor site. This gap is repaired according to a DNA “gap repair” mechanism. This process is simulated as two independent repair processes beginning simultaneously at each side and progressing towards each other. At each position, the processes restore the symbol present in the excised copy. But the repair processes can also make errors.

- The repair process can abort on one side with a probability \( d \). If it aborts at both sides, then an element with a deletion covering the unrepaired region appears. The symbol “−” is used to indicate a deleted position. An element is restored at the donor site if at least one repair process reaches the other on the sequence, even if one has aborted.

- Mutations can occur with a probability \( e \). The original symbol is randomly replaced by another. Obviously, no mutation event can occur at a deleted position: a “−” symbol only appears by abortive gap repair and then cannot be changed to another.

According to the status of the four functional regions, different classes of copies with different properties can occur.

- **Active copies**: If all regions contain at least one “1” or one “*”, the transposase is functional and the copy is active. Its Rec and Endo domains are able to bind and cleave a particular copy.

- **Regulator copies**: If only the Endo domain is devoid of “1”’s, the encoded transposase can bind, but cannot cleave the element. When it binds, all functional transposases are denied access and the copy acts as a regulator.

- **Inert copies**: If the Rec domain is devoid of “1”’s and “*”’s, the transposase is unable to bind to any element and transposes it: the transposase is inert.

- **Immobile copies**: Finally, if one transposase target site is totally deleted (only the “*” symbol), no transposase can bind at this site and thus no cleavage can occur: this copy is immobile.

**IMPLEMENTATION**

This model has been implemented in the simulation program called GENOOM (Quesneville & Anxolabéhère, 1997a). Based on an object paradigm, this program considers and represents each entity of the system to be modeled as a digital object. For speed and memory purposes, the four regions of the sequence were limited to four positions, allowing us to represent a TE copy with a computer word of 32 bits. Populations can be initialized with different kinds of copies and are simulated with the model. All probability tests are performed with a classical Monte-Carlo procedure.

**Results**

**EMERGENCE OF A CLASS II TRANPOSABLE ELEMENT**

With this minimal TE model, the modalities of TE emergence from an endonuclease gene have been explored. The starting population is composed of individuals carrying each one copy of an endonuclease gene [Fig. 1(c)], two chromosomes, and 100 sites per chromatid. At first, this sequence is immobile because the transposase cannot recognize the target site, but recognition could occur by mutation. Four sets of simulation parameters have been tested (Table 2). In addition, for each set, four gap repair error conditions are explored: no error, deletions errors only, mutation errors only, and a mixed errors condition where both deletions and mutations occur.
The values used for the repair errors, are respectively, 0.05 and 0.01 per position and per repair event for deletion and mutation errors. Each simulation condition is repeated 10 times over 1000 generations.

Figure 2 shows that with all parameter sets, the number of copies can increase, meaning that the endonuclease gene has mutated and that the endonuclease recognizes target sites located on each side of the gene. Once the gene has become mobile, a regulation system emerges progressively, as shown by the decrease of the transposition rate several generations later. Thus, the gene has become a true transposable element.

But TE invasion dynamics shows differences according to the type of repair errors made. Repair errors make TEs emerge and invade but also be regulated more rapidly. Consequently, the copy number at 1000 generations is generally less than when no error occurs in the gap repair process. The “no error” condition can exhibit regulation only when mutations are frequent (sets A and C), because regulator copies can only appear during chromosomal replication by mutation. Strong selective pressure (sets C and D) associated with the “mutation error” condition is the most invasive situation and gives more copies than mixed or deletion errors conditions.

### EVOLUTION OF THE SEQUENCE

Simulation results obtained with parameter set A, where mutation and deletion repair errors could both occur, were analysed in order to understand the sequence evolution of TEs. Data were collected for copies present in populations at generation 1000 and accumulated over the 10 repeats.

The transposase target sites, *IdL* and *IdR*, are initialized with the string “0000” in the endonuclease gene. “1”s and “*”s appear by mutation with the same probability. The accumulation of one of these symbols can be interpreted in terms of selective advantage. Figure 3 shows less “1”s than “*”s in the *IdL* and *IdR* regions, whatever the copy type. Thus, copies tend to be recognized by several different transposase Rec domains: the target sites become generalized in their ability to be recognized. This tendency is stronger for regulator copies. This points out the importance of mobility for TEs, which is deleterious for the host and hence counter-selected, but also allows the rapid multiplication of copies which in turn benefits the TE. From these two antagonistic effects emerges the complexity of the TE behavior and specific consequences for their sequence evolution arise (this point will be discussed further).

The transposase/endonuclease recognition domain, called *Rec* domain, is also initialized in the endonuclease gene with the sequence “0000”. “1”s and “*”s also appear by mutation. Active copies show significantly less “*”s than “1”s, respectively 0.46 and 0.54 ($\chi^2$ test performed on the 339202 “1” and “*” symbols counted over all 10 repeats at generation 1000: $\chi^2 = 2245.22; df = 1; p$-value $< 2.2 \times 10^{-16}$). Due to the deleterious effect of active transposases, they are selected to recognize less copies so that they induce less mobility. But regulator copies encoding an inactive transposase that acts as a regulator present more “*”s and “1”s, 0.54 and 0.46, respectively ($\chi^2$ test performed on the 287134 “1” and “*” symbols counted over all 10 repeats at generation 1000: $\chi^2 = 1307.776; df = 1; p$-value $< 2.2 \times 10^{-16}$). They tend to be selected to recognize many different copies. Indeed, if they are able to block many different active transposases by occupying their fixation sites, they reduce the overall copy mobility and are selectively favored.

### DELETED COPIES

Figure 4 shows the overall distribution of deletions. The asymmetrical feature of this distribution can be again interpreted in terms of selective advantage. Indeed, copies that are defective in the *Endo* domain are potential regulator copies; they are positively selected and thus are abundant. Due to their high frequency among the copies, the global distribution shows a distribution skewed toward the *IdR* side, over the *Endo* domain positions.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sets of simulation parameters</strong></td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>Mutation rate ($m$)</td>
</tr>
<tr>
<td>Selective impact ($s$)</td>
</tr>
</tbody>
</table>
FIG. 2. Dynamics of TE emergence from an endonuclease gene. Four sets of simulation parameters are used: A, B, C and D (Table 2). In addition, for each set, four gap repair error conditions are explored: no error, deletion errors only, mutation errors only, and both deletion and mutation errors. The values used for the repair errors are 0.05 and 0.01 per position for deletion and mutation errors, respectively. Each simulation condition is repeated 10 times over 1000 generations. (–––) no error; (––––) deletion errors; (––) mutation errors; (-----) mixed errors.

Table 3 shows the study of the target site deletions of immobile copies. A net excess in deletions in the \textit{IdR} side can be observed ($\chi^2$ test performed on the target site deletion counts over all 10 repeats at generation 1000: $\chi^2 = 1909.068$; df = 1; $p$-value < $2.2 \times 10^{-16}$). This can be explained by successive deletion events: since copies deleted in the \textit{Endo} region are abundant, a new defective element is more likely to appear from them, therefore new deletions occur more frequently in \textit{IdR} than in \textit{IdL}.

Differences in symbol frequencies between \textit{IdL} and \textit{IdR} regions are induced by this asymmetry in the distribution of deletions. The frequencies of “0”s in \textit{IdL} and \textit{IdR} in all copies are, respectively, 0.51 and 0.53 ($\chi^2$ performed on the “0” and “non-0” symbols counted over all 10 repeats at generation 1000: $\chi^2 = 4238.136$; df = 1; $p$-value < $2.2 \times 10^{-16}$). Obviously, this reflects the fact that the more a region is frequently deleted, the less its sequence can mutate. This result could explain the higher frequency of “0”s in the \textit{Rec} region compared to \textit{IdL} and \textit{IdR} (Fig. 3). Another difference is that the frequency ratio of “*”s over “1”s is greater in the \textit{IdR} region (2.22 and 3.44 for \textit{IdL} and \textit{IdR}, respectively). The \textit{IdR} region being
generally more partially deleted, the selective pressure in favor of the generalization tendency described before is here stronger because of the few positions still available.

**Discussion**

**LIMITS OF THE APPROACH**

The purpose of this work was to show what emergent behavior is “mechanistically” possible given the structure and the properties of class II TE. The encoding of the four TE regions only models functional properties and not structural characteristics such as a DNA sequence. Consequently, no relation between the size of the region (four positions) and the length of these region on the DNA sequence could be conjectured. Such a question is beyond the outcomes of this model. Moreover, the implementation of the model is constrained by the speed and memory requirement of the simulations. In order to obtain a dynamic in reasonable simulation time, mutations appear to occur more frequently than it would be expected in real life. Thus, only qualitative outcomes could be discussed. Thus, it is difficult from the dynamics obtained by simulation to deduce the time needed in real life for such dynamics.

This model has to be considered as a null hypothesis. It produces results according to our knowledge of class II TEs structure and transposition mechanisms. Discrepancies from real data about a TE family must be interpreted in terms of additional features to be added to the model, i.e. our knowledge of that TE. These
FIG. 4. Distribution of deletions over all copies. These charts indicate the frequency with which each position is found to be deleted in the copies. They are calculated from the population repeats at generation 1000, obtained with parameter set A with both deletion and mutation errors.

TABLE 3
Target site deletion frequencies of the 441,292 immobile copies from the 10 population repeats at generation 1000, obtained with parameter set A with both deletion and mutation errors. The three classes of immobile copies called here “IdL”, “IdR”, and “IdL and IdR”, are, respectively, copies deleted in the IdL, IdR, and both IdL and IdR regions.

<table>
<thead>
<tr>
<th>IdL</th>
<th>IdR</th>
<th>IdL and IdR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>0.38</td>
<td>0.29</td>
</tr>
</tbody>
</table>

features are very interesting to point out, because they highlight the singularities of that TE.

SOURCE OF THE BEHAVIOR COMPLEXITY

The complexity of the TE behavior results from the opposition of two driving forces acting on the copy mobility. The first counter-selects mobility because of its deleterious impact on host fitness. This is a “mobility disadvantage” for the benefit of the host. The second selects mobility as a rapid multiplication mechanism of TE copies. This is a “mobility advantage” for the benefit of TEs, visible during the initial phase of TE emergence. When the main driving force is the “mobility advantage”, the transposition rate and copy number increase. But when the mobility becomes too high, then it is counter-selected and the “mobility disadvantage” becomes the main driving force. As a result, a mobility control appears from regulator copies. The outcome of this opposition is a dynamic equilibrium state.

SEQUENCE EVOLUTION

The antagonism between these two selective pressures gives rise to the observed heterogeneity among the TE regions but also between the different copy types. Transposase target sites (IdL and IdR regions) increase their capacity to bind different transposases. The target site recognition domain of the transposase (the Rec domain) evolves differently according to the type of copy. In autonomous copies, recognition ability is reduced, while in non-autonomous copies that produce a truncated transposase acting as a competitive inhibitor, the capacity to recognize different target sites is increased.
Deletions enforce heterogeneity by introducing asymmetry in the evolution of sequences. The distribution of the deletions in the sequences is not centered on the middle of the copies as the gap repair mechanism could suggest, but it is shifted to the “endonuclease region”. Consequently, immobile copies, which are deleted at least in one of the two transposase target sites, are more often deleted in the IdR region close to the “endonuclease region”.

But the deletion distribution also affects the capacity of sequences to evolve. Since central regions are less represented, fewer mutations occur in these regions, and the speed of symbol changes may vary across the sequence. Consequently, the two “target sites” may evolve more rapidly than the other regions. Hence, the IdR region close to the “endonuclease domain” would evolve more slowly than the IdL one, because of the shift in the deletions distribution.

**ORIGIN OF CLASS II TEs**

Our results show that an endonuclease origin of class II TEs is compatible with their dynamics and produces copy classes as we know them for real class II TEs. The similarity of transposase to site-specific endonucleases, added to the results presented in this paper, suggest that class II TEs can derive from endonuclease genes. The chemical similarity of reactions and the conserved D,D(35)E amino acid motif indicate that TE transposases belong to a super-family of recombinases that execute metal-ion-dependent polynucleotidyl transferase reactions (Craig, 1995).

The origin of class II TEs has to be found probably in recombinase genes.

The shorter the target site of an endonuclease, the more frequent the cleavage sites are in the genome. Thus, short target site endonucleases are more deleterious than long target site ones because they cut in the genome more often. Consequently, short target site endonucleases would be rare in eukaryote genomes. The scenario presented here starts from an already functional endonuclease gene which is already supported by the hosts. The size of the endonuclease target site is a key factor conditioning the rapidity of the emergence. The bigger the target site is, more time is required for the endonuclease gene to become mobile. Hence because long target site endonucleases would be more frequent, the emergence of a mobility would be rare and have arisen only few times in the evolutionary history.

Our results show that the gap repair process plays an important role in the invasion dynamics. It has already been shown that it could explain the invasion success of the P transposable element after an horizontal transfer (Quesneville & Anxolabéhère, 1997b, 1998). Here, it appears to condition the emergence success of class II TEs. It is when deletion errors are made that regulation appears most rapidly. Consequently, less copies invade than when only mutations or no errors occur. When the gap repair process generates mutation errors, the invasion success is increased.

Differences in the efficiency of gap repair processes in species may account for the presence of class II TEs. The absence of DNA transposable elements in the Saccharomyces cerevisiae genome could be the effect of its efficient DNA gap repair system, as shown by the “no error” simulation condition. In this case, the regulator copies can appear only by mutation during chromosomal replication. The selective pressure against the TE invasion is then maximal, since the TE family faced with the frequent deleterious transposition events must generate regulatory copies early enough to prevent its elimination from the host genome.

We thank Caroline Lemerle and Olivier Andrieu for their assistance and their helpful comments in the preparation of this manuscript. We thank also the anonymous referees for their helpful comments. This work was supported by the Centre National de la Recherche Scientifique (Institut Jacques Monod: UMR 9922, and by the programme génome, by Paris 6 and Paris 7 universities.

**REFERENCES**


