# Model-based Identification of Helitrons Results in a New Classification of Their Families in Arabidopsis thaliana

Sébastien Tempel, Jacques Nicolas, Abdelhak El Amrani, Ivan Couée

**Title:** Model-based Identification of Helitrons Results in a New Classification of Their Families in *Arabidopsis thaliana*

**Keywords:** Bioinformatics, sequence analysis, genome dynamics, syntactical modeling, transposable elements, helitron, chimera, combinatorial optimization.

**Authors:** Sébastien Tempel[1,2,*], Jacques Nicolas[1], Abdelhak El Amrani[2] and Ivan Couée[2]

**Affiliations:**

[1] IRISA-INRIA, Campus de Beaulieu Bât 12,

35042 Rennes cedex, France

[2] CNRS, Université de Rennes 1, UMR 6553 Ecobio,

Campus de Beaulieu Bât 14A,

35042 Rennes cedex, France

**Address for Correspondence:**

Sébastien Tempel

IRISA-INRIA, Bureau A113, Campus de Beaulieu Bât 12,

35042 Rennes cedex, France

Phone: +33 (0) 2 99 84 73 23

Fax: +33 (0) 2 99 84 71 71

Email: sebastien.tempel@irisa.fr

ABSTRACT

Helitrons are a class of prolific transposable elements in the *Arabidopsis thaliana* genome. Although 37 families were identified after the recent discovery of helitrons, no systematic classification is available because of the high variability of helitronic sequences. Since transposition proteins are assumed to interact with helitron termini, a helitron model was formalized based on terminus characterization in order to carry out an exhaustive analysis of all possible combinations of the pairs of termini present. This combinatorics approach resulted in the discovery of a number of new helitron elements corresponding to termini associations from distinct previously-described helitron families. The occurrence matrix of termini combinations yielded a structure that revealed clusters of helitron families. This sheds light on the history of their invasion of the *Arabidopsis thaliana* genome.

INTRODUCTION

Transposable Elements (TEs) move or are copied from one genomic location to another [Feschotte 2002]. TEs have been described in all eukaryotic and prokaryotic genomes [Kidwell 2001]. They are characterized and classified on the basis of terminal or subterminal remarkable structures or their protein-coding capacity. Class I elements move via an RNA intermediate and encode a reverse transcriptase. Class II elements or DNA transposons seem to move via “cut-and-paste” mechanisms where the DNA element itself is the mobile intermediate. TE copies that do not show any coding capacity are considered to be non-autonomous elements that require transposition proteins from autonomous elements for transposition [Feschotte 2000].

A new family of DNA eucaryotic transposons, called helitrons, has been described recently in plants and other eukaryotes [Kapitonov 2001, Feschotte 2001]. Like *Geminivirus*, autonomous helitrons code for ssDNA-binding replication protein A ((RPA)-like protein A) and helicase, which are involved in transposition [Kapitonov 2001, Feschotte 2001, Guiterrez 1999, Iftode 1999]. Helitrons are characterized by typical terminal and subterminal structures: a TC 5' terminus, a CTAG 3' terminus, and a 3’ subterminal short hairpin structure [Kapitonov 2001, Eckardt 2003]. Non-autonomous helitrons are characterized by large mutations, indels of the internal sequence of autonomous helitrons, keeping in common just the typical terminal and subterminal structures. They have received the collective name of AtREP in the model plant *Arabidopsis thaliana* [Kapitonov 2001]. The name Helitron is usually ascribed to autonomous helitron families (Helitron 1, 2, 3, 4 and 5) and to long non-autonomous helitrons, such as Helitron y1A, y1B, y1C, and y1D [Kapitonov 2001]. Autonomous and non-autonomous helitrons have been classified according to the homologies detected in their sequence in the Repbase database [Jurka 2005].

Five autonomous and 32 non-autonomous helitron families have been described in the *Arabidopsis thaliana* genome [Kapitonov 2001, Jurka 2005]. Multiple alignment of consensus sequences of non-autonomous helitron families and visualization by DomainOrganizer [Tempel 2006] clearly show that helitronic extremities and a large subterminal sequence are similar in all families (Supplementary Material 1). Since transposition proteins are thought to recognize the termini of non-autonomous transposable elements [Kapitonov 2001, Feschotte 2001, Jiang 2004], common terminal and subterminal structures are likely to be characteristic of helitrons that depend on the same transposition proteins and share similar dynamics in copy amplification.

The classification and characterization of helitrons was therefore analyzed on the basis of their terminal and subterminal sequences in the whole genome of *Arabidopsis thaliana*. This approach was compared with the classification based on whole sequences of the Repbase database [Jurka 2005].

The systematic study of all possible pairs of 5' and 3' termini was shown to provide a structured distribution of occurrences, which this paper proposes to use as the basis for a new classification.

## MATERIAL AND METHODS

### *Genomic data*

The 03/17/2004 version of the *Arabidopsis thaliana* genome sequence was obtained from the TAIR website (www.arabidopsis.org). The initial set of helitron sequences was obtained from the Repbase database (www.girinst.org/repbase/index.html) [Jurka 2005].

*RepeatMasker program*

For each family of Arabidopsis present in Repbase, the number of occurrences of helitrons with one terminus, two termini and no termini was computed. The number of sequences showing a size similar (+/- 10 %) to the size of the consensus sequence present in Repbase was also calculated. This was achieved using RepeatMasker version open-3.1.6 with default parameters. The software was obtained from the RepeatMasker web site (www.RepeatMasker.org). The library associated with RepeatMasker was the latest version of the Repbase library for RepeatMasker (www.girinst.org/repbase/index.html).

*Syntactical model of helitron families*

The helitron model described in the literature [Kapitonov 2001] (TC in 5' and CTAG with subterminal hairpin in 3') is a pattern that is too general to provide precise recognition of helitronic families. In order to create a model characterized in the same way by two helitronic termini with a variable gap in between, the study first investigated the optimal size of termini required to distinguish them from the rest of the genome. Consensus sequences of helitronic termini were extracted from Repbase [Jurka 2005]. AtREP16, 17, 18 and 19 were excluded from this set, since these families do not have the characteristic termini of helitrons [Kapitonov 2001]. The size of the gap was set according to the observed maximum size of known helitrons [Kapitonov 2001, Jurka 2005]. The optimal size of helitronic termini and the maximum number of substitutions were determined by minimizing the difference between the number of matching sequences and the corresponding data from Repbase [Kapitonov 2001; Jurka 2005].

*Exhaustive search of the terminus-based helitron model*

Occurrences of the termini and the helitron model were parsed using STAN [Nicolas 2005]. STAN recognizes a subset of SVG (String Variable Grammars) [Dong 1994, Searls 2002] and can search complex biological patterns such as palindromes or repeats in genomes.

The presence of subterminal hairpins (6 to 8 nucleotides), described in previous models [Kapitonov 2001, Feschotte 2001, Eckardt 2003], was searched in all detected helitronic sequences using STAN [Nicolas 2005]. The chosen model represents a 6- to 8-bp hairpin with a 4- to 5-nucleotide loop. Using STAN syntax, this is written: X:[6,8]-x(4,5)-~X.

*Exhaustive search and analysis on helitronic termini combinations*

"Left" and "right" refer to the 5’ and 3’ termini of a given family of helitrons. LEFT and RIGHT are respectively defined as the complete set of 5' termini and the complete set of 3' termini of all helitronic families extracted from Repbase [Jurka 2005]. For each possible pair of termini ($left_i$, $right_j$) $\in$ LEFT X RIGHT, a grammar was produced and submitted to STAN and the genome of *Arabidopsis thaliana* was parsed. This resulted in a frequency matrix of hits on LEFT X RIGHT. A cell ($left_i$, $right_j$) of this matrix contains the number of instances of the models starting with $left_i$ and ending at a suitable distance with $right_j$. This definition applies to embedded, overlapping and chimeric helitrons (created by combining two distinct sequences).

*Aggregating helitronic extremities and pairs of extremities*

The LEFT and RIGHT sets are quite large in size, as a result of the fine extremity patterns used. In order to rationalize the choice of patterns, the number was first reduced by forming equivalence classes. This was achieved based on the extent of termini (set of occurrences) in the

genome. More precisely, let $f_{left}^{ij}$ ($f_{right}^{ij}$) denote the frequency of sequences covered by the $left_i$ ($right_i$) pattern and not covered by the $left_j$ ($right_j$) pattern. A standard hierarchical classification algorithm was applied, starting with a set of singletons corresponding to the set of termini, and at each step aggregating the classes at a minimum distance. The distance between two classes $c_1$ and $c_2$ is defined as:

$$d(c_1,c_2) = Min_{x \in (c_1 \cup c_2)} \sum_{y \in (c_1 \cup c_2) - \{x\}} f^{yx}$$

The argument x used to minimize the equation represents the class $c_1 \cup c_2$. Aggregations were retained when the distance was less than 10 % of the number of instances covered by $c_1 \cup c_2$.

*Rearrangement of rows and columns in the matrix of occurrences*

The highest values of occurrences of termini combinations were assumed to reflect the genuine associations that emerged at the origin of the families. To trace back these founding combinations, the iterative optimization algorithm of Munkres was used [Munkres 1957; Bourgeois 1971]. The matrix was sorted to show these preferential associations on the diagonal.

*Exhaustive study of autonomous helitrons*

For each helitron sequence detected by STAN models, the study searched for ORFs using GENSCAN [Burge 1997], followed by BLASTP [Altschul 1997] to identify them.

RESULTS

*Syntactical helitron model, identification and comparison using RepeatMasker*

Analysis showed that termini as long as 36 bp were necessary and sufficient to define and retrieve a given family of helitrons from Repbase. These 36-bp structures encompass a larger

region than the canonical TC at the 5’ end and include the subterminal hairpin at the 3’ end (Figure 2). Alignments in most cases showed a certain level of polymorphism in these 36-bp sequences and using exact termini sequences was insufficient. For example, searching for AtREP3 with exact termini yielded only 13 occurrences, a much lower value than the 150 occurrences reported in the literature [Kapitonov 2001]. Therefore, as transposable elements are known to accumulate mutations between generations, a substitution rate of 25 % was introduced in SVG models. Using 36-bp termini and 9 errors, all occurrences for families in Repbase were detected. For instance, searching for AtREP3 with 9 errors returned 141 occurrences, which was in line with the number of occurrences given by Repbase.

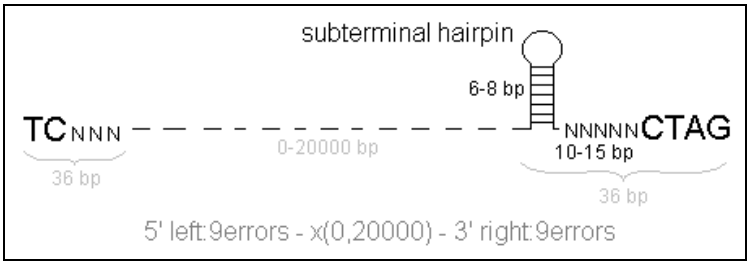


Figure 2: Relationship between current biological knowledge and our syntactic model. Data available in the literature is in black, knowledge obtained from preliminary studies is in grey and the bottom line represents the model with two termini of 36 nucleotides, a threshold error of 25% and a variable gap with a maximum length of 20,000 bp.

The 141 sequences were aligned using the AtREP3 consensus downloaded from Repbase [Jurka 2005]. Multiple alignment showed that most occurrences of AtREP3 were similar to the AtREP3 consensus. The other occurrences show a large deletion of the 5’ subterminal sequence. The result confirms the relevance of this new helitron model. The corresponding helitron model was written as follows in the formalism described in Materials and Methods: $left_i$:9 - x(0,20000) - $right_j$:9.

The syntactical method used in this study was compared with the RepeatMasker identification for all known helitronic families (Figure 3). The method used WU-BLAST to compare the library of transposable elements against query sequences or genomes. In almost all cases STAN detected correctly sized sequences (± 10 % of Repbase consensus) more efficiently. In contrast, RepeatMasker detected a large number of incomplete helitron copies that were significantly smaller than the consensus sequence in a given family (Table 1). Most of these sequences lacked the typical 5' and 3' termini. Moreover, the average number of occurrences detected by STAN was greater than the number of occurrences of the corresponding consensus sequence. STAN is capable of detecting certain helitrons that include other transposons in their internal sequences, such as in the AtREP21 family [Tempel 2006]. A comparison was also made between the total number of helitrons detected using both methods (Figure 3). Except for 37 sequences, which display all the helitron characteristics, all the sequences detected by STAN were entirely or partially detected by RepeatMasker. On the contrary, most sequences detected by RepeatMasker were not detected by STAN. This is due to the fact that more than 80 % of the sequences detected by RepeatMasker are partial helitrons (Figure 3).

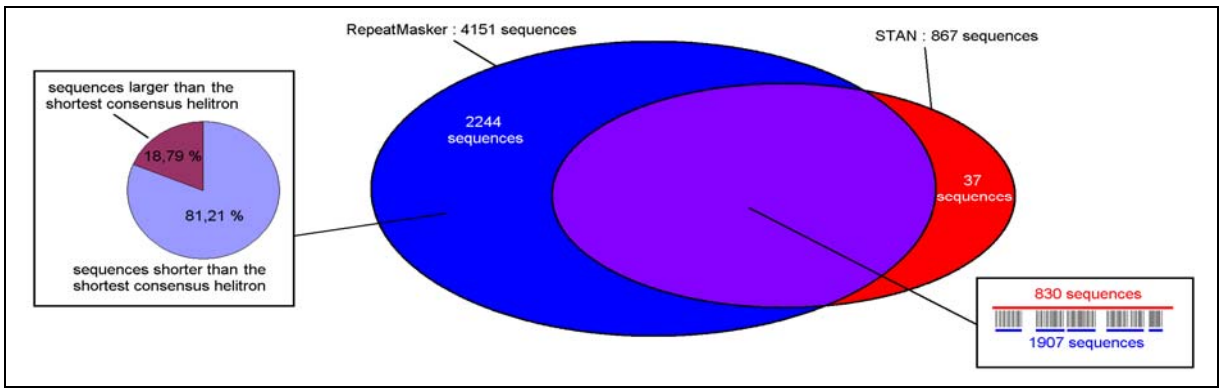


Figure 3: Comparison of sequences detected by RepeatMasker and STAN. Sequences detected by both methods are in purple, those detected only by RepeatMasker are in blue and those by STAN in red. The right insert details the

number of common sequences detected by RepeatMasker (in blue) and STAN (in red). The number of blue sequences is more than two times greater, since STAN sequences generally cover several sequences detected by RepeatMasker. The left insert shows the size distribution of sequences detected by RepeatMasker and not detected by STAN in relation to the size of the shortest consensus helitron (564 bp) [Kapitonov 2001].

*Updating the helitron model*

The search for subterminal hairpins in helitronic 3' extremities showed a low proportion of exact palindromes. There were only 467 sequences out of a total of 867 that contained subterminal hairpins. When the search checked only for hairpin structures (i.e. for palindromes) without considering the underlying sequence, the analyzer detected non-helitronic sequences (data not shown). Given these results, the choice was made to not require a hairpin in the 3' termini for the purposes of this paper.

*Genome-wide analysis of termini occurrences: evidence of truncated helitrons*

Genome-wide analysis of the distribution of each type of helitron termini (Table 2 and 3) showed that extremities of defined families were unexpectedly clustered with extremities of other families. For example, 5' and 3' extremities of AtREP2 and AtREP2A or extremities of AtREP6, 7, 8, 9 were always associated. Certain extremities of helitron families, such as those of AtREPX1, never co-occur with any other extremities. Moreover, the clusters obtained on the 3' extremities do not always correspond to clusters on 5' extremities. For example the 5' terminus of AtREP3 was associated with the 5' terminus of AtREP20 (Table 2), while the 3' terminus of AtREP3 was associated with the 3' terminus of AtREP11 (Table 3). Clustering the extremities according to the methods developed in Material and Methods resulted in the identification of 23 (1 to 23) types of 5' (Table 2) and 23 (a to w) types of 3' extremities (Table 3).

| Name | Helitron family name [Kapitonov 2001] | Sequences | Occurrences | Lost (%) |
|---|---|---|---|---|
| 1 | HELITRON1 | TCTACATATACATTTTTGGGGACGATTTGTGTCGAA | 84 | 0,00 |
| 2 | HELITRON2, Y2 (HelitronY2) | TCTACACTATTATTTGAGACGTACGTTAAGTGATTC | 32 | 12,50 |
| 3 | HELITRON3 | TCTACTTAACATTTTTGAAGTACAAAATAAGGAATT | 44 | 0,00 |
| 4 | HELITRON4 | TCTATACTATTAAGAGAGAATGAGGGACAGAAAATT | 14 | 0,00 |
| 5 | AtREP1, 11 (AtREP11) | TCTACATATACATTTTTGCAGCCATTTTAGCAAATA | 453 | 2,43 |
| 6 | Helitron5, AtREP2, 2A, 11A, 14 (AtREP2) | TCTATATATACATTTTTGCAGCCATTTTGTGAAATA | 488 | 5,33 |
| 7 | HELITRONY1A | TCTACATATACATTTTTGGTAGGGTTTTTGGCCAAA | 86 | 0,00 |
| 8 | HELITRONY1B | TCTACATATACATTTTTGGTAGACTTTTTGGCCAAA | 79 | 0,00 |
| 9 | HELITRONY1C | TCTACATATACATTTTTGGAACGATTTGTGTAGAAA | 189 | 0,00 |
| 10 | HELITRONY1D | TCTACTTAACATTTTTGAAGTACATTTTAAGGAATC | 45 | 0,00 |
| 11 | HELITRONY1E | TCGACATTTTAAATCAAATATTTACCTAACAAAAAT | 21 | 0,00 |
| 12 | HELITRONY3 | TCTTATATAATAAGAGAAGGTTTTTTCTAACTTTGC | 38 | 0,00 |
| 13 | HELITRONY3A | TCTACTTAACATTTTTGAAGTACAAAATAAGGATTT | 47 | 0,00 |
| 14 | AtREP3, 20 (AtREP3) | TCCTACTATATTATTTGGGAAGTACATTTTAAATGT | 220 | 0,00 |
| 15 | AtREP4 | TCTATATTATTAATTGGGAAGCATTGTGTGAACTAC | 60 | 0,00 |
| 16 | AtREP5 | TCTATATATACATTTTTGGAGGGGATTTTTGAAGAT | 140 | 0,00 |
| 17 | AtREP6, 7, 8, 9, 13 (AtREP6) | TCATATATATGAAAGTTGGCCAACTCTCTCCATATA | 241 | 6,64 |
| 18 | AtREP10, 10A, 10B (AtREP10A) | TCTTATTATATAAAGTATGGTTTTCAAAGTTACTAA | 255 | 1,96 |
| 19 | AtREP10C, 10D (AtREP10D) | TCTTATTATATAAAGTATGGTTTTTAAAATTACTAA | 264 | 3,41 |
| 20 | AtREP12 | TCTAAATATACTAAATCAGCAGTCACTTTTCCAATA | 35 | 0,00 |
| 21 | AtREP15 | TCTATACTATTAATTGGGAATCATTGAAAATAAACA | 35 | 0,00 |
| 22 | AtREP21 | TCCCTTTATTATTAAAAGGGAAGTACAAATTGAAAT | 82 | 0,00 |
| 23 | AtREPX1 | TCAACACCATAAAAACACTAAAAGTCTCTCCTGTGC | 20 | 0,00 |
| | | Total | 2972 | 2,39 |

Table 2: Number of occurrences of 5' termini for each helitron family in *Arabidopsis thaliana*. The first column labels 5' extremities. The second column indicates the previous names of the 5' extremities [Kapitonov 2001], and in parentheses gives the representative for this cluster, while the fourth column indicates its number of occurrences. The last column shows the loss of occurrences when reducing a cluster to its representative.

| Name | Helitron family name [Kapitonov 2001] | Sequences | Occurrences | lost (%) |
|---|---|---|---|---|
| a | HELITRON1, Y1C (HELITRONY1C) | TAATCAACCCGCGGTGTACCGAGGGTCAATATCTAG | 336 | 2,38 |
| b | HELITRON2, Y2 (HELITRONY2) | CCAAAGATACCGTGCGTAGCACGGGTACTGACCTAG | 25 | 0,00 |
| c | HELITRON3 | AAAAAATCTCGCGGTGTACCACGGGTCATATCCTAG | 144 | 0,00 |
| d | HELITRON4 | TTTAACACCCGGGCGAAGCACGGGTTATCAATCTAG | 14 | 0,00 |
| e | HELITRON5 | AACTATCCCTGCGGTGTACCGAGGGTCAAAATCTAG | 449 | 0,00 |
| f | HELITRONY1A | ATATCCACCCGCGGTACACCGCGGGTCAATATCTAG | 429 | 0,00 |
| g | HELITRONY1B | ATATCAACCCGCGGTGTACCGCGGGTCAATCTCTAG | 472 | 0,00 |
| h | HELITRONY1D | TATTTAACCCGCGGTATACCGCGGGTCATATCCTAG | 228 | 0,00 |
| i | HelitronY1E, AtREP11A (AtREP11A) | ATATAGCCCCGCGGTATACCGCGGGTTAAATTCTAG | 580 | 7,59 |
| j | HELITRONY3 | TATATAAAATCCACGCATCGCGTGGACAACTTCTAG | 11 | 0,00 |
| k | HELITRONY3A | CAAAAATCCTGCGGTGTACCGCGGGTCATATCCTAG | 86 | 0,00 |
| l | AtREP1, 2, 2A (AtREP2A) | AATTGTCTCCGCGGTATACCGCGGGTTAAAATCTAG | 577 | 7,63 |
| m | AtREP3, 11 (AtREP3) | AAATCGTCCCGCGGTATACCGCGGGTTAAAATCTAG | 672 | 5,51 |
| n | AtREP4, 15 (AtREP15) | ATGAAATCCCGCACGTACGTGCGGGTCAGGATCTAG | 71 | 2,82 |
| o | AtREP5 | ATATAGGCCCGCGGTATACCGCGGGTTAATATCTAG | 597 | 0,00 |
| p | AtREP6, 7, 8, 9 (AtREP6) | AAAACAAACCCACGCGTAGCGTGGGTACTCATCTAG | 126 | 7,14 |
| q | AtREP10, 10A, 10B, 10C, 10D (AtREP10C) | ATCTTACACCCGCCTATTTAGGCGGGCCTTATCTAG | 159 | 6,29 |
| r | AtREP12 | CAATGTTGTCCCTGCATAGCAGGGGCCAAATGCTAG | 8 | 0,00 |
| s | AtREP13 | CAAAAAAACCGCGGCGTAGCGCGGGTGATCACCTAG | 74 | 0,00 |
| t | AtREP14 | TATTTGCCCCGTGGTGTACCACGGGTTAAATCCTAG | 232 | 0,00 |
| u | AtREP20 | AAATGCCCGTGCTGTATAGCACGGGTTATGATCTAG | 64 | 0,00 |
| v | AtREP21 | CCGATTGCCCGCGGTAAACCGCGGGTTAAAACCTAG | 189 | 0,00 |
| w | AtREPX1 | AAAAAAGCCCGCGGCTTGGCGCGGGTTTCGCCCTAG | 20 | 0,00 |
| | | Total | 5563 | 2,77 |

Table 3: Number of occurrences of 3' termini for each helitron family in *Arabidopsis thaliana*. The first column labels 3' extremities. The second column indicates the previous names of the 3' extremities [Kapitonov 2001], and in parentheses gives the representative for this cluster, while the fourth column indicates its number of occurrences. The last column shows the loss of occurrences when reducing a cluster to its representative.

Surprisingly enough, except for some families such as Helitron2, 4, Y2, AtREP4 and X1, occurrences of 5' termini and 3' termini did not share a one-to-one relationship. For instance, the 5' terminus of AtREP20 was three times more frequent than that of the corresponding 3' terminus, and conversely, the 3' terminus of AtREP3 shows a number of occurrences three times higher than that of the corresponding 5' terminus (Table 2 and 3). Analysis of this discrepancy revealed that many of these 3' AtREP3 termini were not associated with any helitron-like 5' termini, and that the associated sequence corresponded to a truncated AtREP3.

*Genome-wide analysis of helitronic termini combinations*

All possible pairs in the 5' and 3' LEFT and RIGHT termini sets were searched by STAN [Nicolas 2005] according to the model shown in Figure 2. The resulting number of occurrences is given in Figure 4. It shows the matrix structure of the observed occurrences after reorganizing the rows and columns and associating the termini (see Material and Methods). All of the previously-known families of helitrons were detected and retrieved at the expected level of occurrence. In general, a high correlation was found between the 5' and 3' termini of a given family. For instance, the 5' terminus of AtREP4 was mainly associated with the AtREP4 3' terminus (combination 15-n).

| 5' \ 3' | j | w | s | p | b | n | q | r | d | u | t | v | e | o,i,l,m | f | g | a | h | k | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 3 | 3 | 3 | 5 | 5 | 1 | 1 | 0 |
| 23 | 0 | 14 | 4 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 17 | 0 | 1 | 62 | 106 | 1 | 4 | 9 | 0 | 0 | 4 | 13 | 10 | 17 | 26 | 13 | 21 | 14 | 15 | 15 | 2 |
| 2 | 0 | 0 | 0 | 1 | 17 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 15 | 0 | 0 | 2 | 0 | 0 | 47 | 2 | 1 | 0 | 1 | 1 | 1 | 4 | 14 | 9 | 9 | 6 | 1 | 2 | 3 |
| 21 | 0 | 0 | 2 | 2 | 0 | 20 | 4 | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 4 | 1 | 1 | 1 | 0 | 0 |
| 18,19 | 0 | 1 | 3 | 5 | 1 | 8 | 136 | 0 | 2 | 2 | 13 | 9 | 21 | 27 | 23 | 16 | 8 | 9 | 5 | 1 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 1 | 1 | 0 | 2 | 2 | 1 | 1 | 2 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0 |
| 22 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 0 | 0 | 3 | 53 | 6 | 29 | 10 | 12 | 9 | 13 | 3 | 3 |
| 14 | 2 | 0 | 0 | 1 | 0 | 9 | 6 | 0 | 1 | 52 | 48 | 64 | 117 | 204 | 150 | 137 | 134 | 70 | 27 | 28 |
| 9 | 1 | 1 | 6 | 4 | 1 | 2 | 4 | 1 | 0 | 2 | 64 | 21 | 115 | 136 | 69 | 91 | 68 | 42 | 17 | 19 |
| 5,6 | 1 | 1 | 8 | 5 | 3 | 6 | 6 | 1 | 1 | 14 | 182 | 87 | 323 | 402 | 175 | 189 | 149 | 119 | 27 | 27 |
| 16 | 0 | 1 | 2 | 4 | 0 | 3 | 3 | 0 | 0 | 4 | 32 | 32 | 88 | 111 | 71 | 68 | 54 | 36 | 1 | 4 |
| 1 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 3 | 21 | 11 | 41 | 56 | 43 | 46 | 32 | 23 | 7 | 9 |
| 7 | 0 | 2 | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 5 | 23 | 28 | 41 | 55 | 34 | 11 | 3 | 5 |
| 8 | 0 | 2 | 2 | 1 | 0 | 0 | 4 | 0 | 1 | 0 | 11 | 2 | 22 | 26 | 34 | 52 | 32 | 13 | 6 | 5 |
| 3,10,13 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 12 | 2 | 7 | 10 | 7 | 12 | 13 | 24 | 17 | 17 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 8 | 9 | 7 | 8 | 4 | 2 | 1 | 1 |

| x<10 | x<20 | x<30 | x<50 | x<100 | x<200 | x>200 |
|---|---|---|---|---|---|---|

Figure 4: Frequency matrix of occurrences of all possible pairs of 5’ and 3’ termini corresponding to the model in Figure 2. Each cell is colored according to its value within a 7-grade scale. Each line represents a 5' 36-bp terminus and each column represents a 3' 36-bp terminus as defined in Materials and Methods and in Figure 2. Blue rectangles delimit clusters of superfamilies.

A number of new occurrences were detected, however, thus increasing the estimate of whole helitron sequences in the Arabidopsis genome from 870 copies to 1504 copies (including overlapping helitrons). These new occurrences correspond to previously undetected combinations of helitron termini: for example, 5' terminus number 14 (AtREP3 and 20) was found to be frequently associated with the 3' terminus labeled l (AtREP1, 2 and 2A) (171 occurrences). The internal sequences of this combination were found to consist of domains present in other helitrons combined with new domains, some of these sequences occurring frequently in the genome, thereby indicating that such combinations can be transposed and considered as new helitrons. Certain new combinations do not correspond to new helitrons, however, but rather clusters of termini around known helitrons.

Overall, the distribution pattern of these associations was not at all random, and clearly segregated into clusters of associations. For example, the 3' terminus of AtREP3 (o,i,l,m) was associated 378 times with the 5’ terminus of AtREP1 (number 5), and only 196 times with the 5’ terminus of AtREP3 (number 14 in Figure 4).

*Organization of helitron clusters suggests various transposition activities*

Four clusters of occurrences can be deduced from the matrix shown in Figure 4. The first cluster (upper left in matrix) corresponds mainly to a group of AtREP or helitron families previously defined in Repbase. Each family has a high number of occurrences. The second cluster (upper right in matrix) is characterized mainly by new combinations of termini that are not described in Repbase. For example, the most frequent model is a new combination of 5’ termini number 18,19 with o,i,l,m 3’ termini. The third cluster (lower left) is characterized by a small number of occurrences for each combination of termini. The last cluster (lower right) corresponds to most occurrences of AtREP and helitron in the *Arabidopsis thaliana* genome (Figure 4). These differences in the number of occurrences, between combinations and between clusters, probably depend on the recognition of autonomous helitrons by transposition proteins.

*Identification of new families of autonomous and non-autonomous helitrons*

Since autonomous helitrons are likely to be required for the transposition of all types of helitrons, whether autonomous or non-autonomous, [Feschotte 2001], the possibility of new autonomous helitrons was therefore checked for using GENSCAN (genes.mit.edu/GENSCAN.html) [Burge 1997] in order to detect ORF sequences, and using BLASTP [Altschul 1997] (www.ncbi.nlm.nih.gov/BLAST/) to identify putative functions of these ORFs. A number of long helitron sequences was found to contain ORFs encoding helicase-

like and/or RPA-like proteins (Figure 5). The presence of helicase-like protein is always associated with RPA-like protein. Multiple alignments show that all of these ORFs correspond to autonomous helitrons discovered by Kapitonov et al. [Kapitonov 2001].

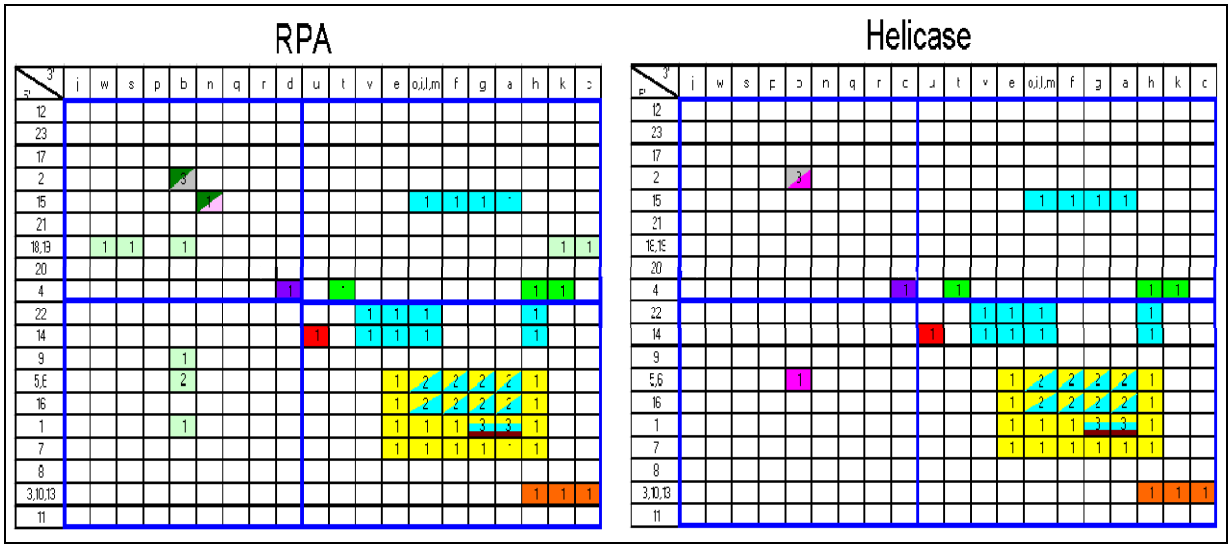


Figure 5: Occurrences of RPA-like and helicase-like encoding ORFs in the helitron sequences for each combination of termini. A different color is chosen for each occurrence.

Most combinations containing RPA-like proteins (40 out of 44 occurrences) and helicase-like proteins (25 out of 32 occurrences) shared the same ORFs with other combinations (Figures 5 and 6). For example, multiple alignment by ClustalW [Thompson 1984] and visualization by DomainRender [Tempel 2006] of autonomous helitrons containing RPA-like and helicase-like protein at position 3200666 to 3210806 in chromosome II, showed multiple combinations of termini at either end of the helitron (Figure 6). Moreover, like autonomous helitrons discovered in bats [Pritham 2007], some helitrons seem to contain an "unknown protein" with RPA-helicase proteins (Figure 6).

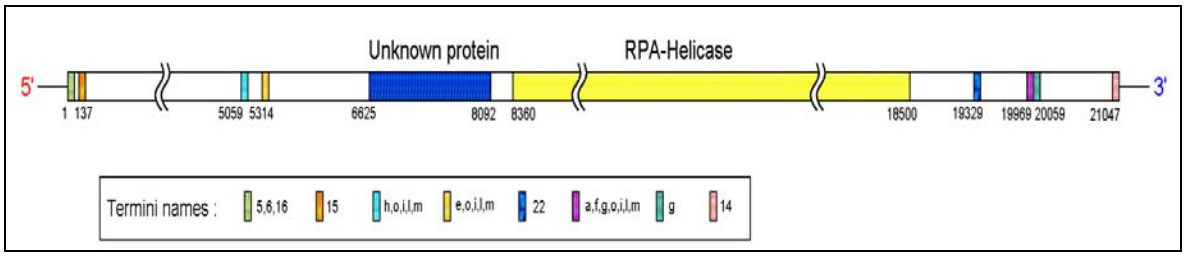


Figure 6: Visualization by DomainRender [Tempel 2006] of multiple termini for one RPA-helicase protein. Red and blue represent the 5' extremity and 3' extremity, respectively. The two ORFs do not have the same orientation (unknown protein: orientation +; RPA-helicase: orientation -).

*New helitron nomenclature*

Figure 4 shows that there are 1369 combinations of termini which could represent nearly 1369 helitronic families in a terminus-based classification. Since multiple combinations of termini were observed at the same location and thus for the same helitron sequence (Figure 6), the study attempted to select the minimum set of termini pairs that corresponds to all of the observed occurrences. The set of occurrences O was first defined, containing the instances of maximum-sized helitrons: an occurrence in O starts with a 5' sequence in the LEFT set, ends with a 3' sequence in the RIGHT set, and is not included in any other occurrence. The termini present in each occurrence are then examined: each element of O was associated with the set of pairs from C = LEFT X RIGHT included in this element. An attempt was then made to solve the associated set covering problem: find the smallest subset of C that covers all elements of O. Since this is an NP-difficult problem, the best to be expected are good heuristic solutions. The standard greedy algorithm [Cormen 2001] scans the elements in C and at each step chooses the termini pair that covers the greatest number of occurrences. It then removes this pair from C and from all occurrences in O where this combination exists. The algorithm iterates until there are no remaining occurrences. A new algorithm was created, which also chooses the pair of termini with

the greatest number of occurrences, while applying one of two alternatives: either keeping this pair, or replacing it recursively to achieve the best coverage of the elements it represents using the remaining pairs. The alternative chosen is the one that leads to the best overall coverage using a minimum number of pairs.

A more precise algorithm is provided in the supplementary material.

The new algorithm returned 44 pairs of termini covering O, thus all the helitronic sequences in the *Arabidopis thaliana* genome (Figure 7). Except for pairs 1_a and 12_j, all of the pairs were directly or indirectly connected to autonomous helitrons. Almost all pairs showing a high level of occurrence had at least one extremity in common with autonomous helitrons. A reasonable number of families showing a significant link between autonomous and non-autonomous helitrons was therefore obtained. Further analysis focused on termini pairs that yielded less than 5 occurrences, while showing no connections to any autonomous helitrons. These pairs seemed to correspond to extremities that degenerated through accumulated mutations (data not shown), and it was therefore decided to leave them out.

| 5' \ 3' | a | j | e | c | f | w | b | u | n | o,i,l,m | t | v | q | g | p | s | h | k | d | r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1(1) | | | | | | | | | | | | | | | | | | | |
| 12 | | 9 | | | | | | | | | | | | | | | | | | |
| 16 | | | 1 | 2 | 1 | | | | | | | | | | | | | | | |
| 9 | | | | | 3 | 1 | | | | | | | | | | | | | | |
| 23 | | | | | | 10 | 1 | | | | | | | | | | | | | |
| 2 | | | | | | | 12(4) | 1 | 1 | | | | | | | | | | | |
| 21 | | | | | | | | | 5 | | | | | | | | | | | |
| 15 | | | | | | | | | 24(2) | 1 | | | | | | | | | | 1 |
| 14 | | | | | | | | 44(1) | 2 | 116 | 1 | | 1 | | | | 1 | | | |
| 5,6 | | | | | | | | | | 284(2) | 13 | 2 | | | | | | | | |
| 22 | | | | 1 | | | | | | | | 40 | 1 | | | | 1 | | | |
| 8 | | | | | | | | | | | | 1 | 2 | | | | | | | |
| 18,19 | | | | | | | | | | | | | 80 | 1 | | | | 1 | | |
| 7 | | | | | | | | | | | | | | 27(1) | | | | | | |
| 17 | | | | | | | | | | | | | | 2 | 85 | 38(2) | | | | |
| 3,10,13 | | | | | 1 | | | | | | | | | | 1 | | 15(1) | | | |
| 4 | | | | | | | | | | | | | | | | | | 1(1) | 2(1) | |
| 20 | | | | | | | | | | | | | | | | | | | | 6 |
| 11 | | | | | | | | | | | | | | | | | | | | |

x<10 x<20 x<30 x<50 x<100 x<200

Figure 7: Matrix of termini combinations that cover all the occurrences of helitron in the *Arabidopsis thaliana* genome. Each cell is colored according to its frequency in a 6-grade scale. Each line represents a 5' 36-bp terminus and each column represents a 3' 36-bp terminus as defined in Materials and Methods and in Figure 3. Numbers in parentheses indicate the number of autonomous helitrons corresponding to a given termini combination.

A new nomenclature has been proposed for the 19 remaining termini combinations (Table 4). The following naming rules were chosen for families: all combinations that contained autonomous helitrons are named "Helitron" followed by a number, and any other combinations are named "AtREP" followed by a number. If the two former extremity names [Kapitonov 2001] are identical and meet the above condition, the former family name has been kept. This nomenclature shows many new autonomous helitronic families (Helitron6, 7, 8, 9 and 10 in Table 4). Nevertheless, they are very similar to autonomous sequences present in Repbase. This suggests that they derive directly from previously known autonomous helitrons.

| Couple | Old name of 5' extremity | Old name of 3' extremity | New Name |
|---|---|---|---|
| 1_a | Helitron1 | Helitron1, Y1C | Helitron1 |
| 2_b | Helitron2, Y2 | Helitron2, Y2 | Helitron2 |
| 3,10,13_h | Helitron3, HelitronY1D, HelitronY3A | HelitronY1D | Helitron3 |
| 4_d | Helitron4 | Helitron4 | Helitron4 |
| 5,6_o,i,l,m | Helitron5, AtREP1, 2, 2A, 11, 11A, 14 | HelitronY1E, AtREP1, 2, 2A, 3, 5, 11, 11A | Helitron5 |
| 7_g | HelitronY1A | HelitronY1B | Helitron6 |
| 4_k | Helitron4 | HelitronY3A | Helitron7 |
| 17_s | AtREP6, 7, 8, 9, 13 | AtREP13 | Helitron8 |
| 14_u | AtREP3, 20 | AtREP20 | Helitron9 |
| 15_n | AtREP4 | AtREP4, 15 | Helitron10 |
| 12_j | HelitronY3 | HelitronY3 | AtREP1 |
| 18,19_q | AtREP10, 10A, 10B, 10C, 10D | AtREP10, 10A, 10B, 10C, 10D | AtREP10 |
| 20_r | AtREP12 | AtREP12 | AtREP12 |
| 5,6_t | Helitron5, AtREP1, 2, 2A, 11, 11A, 14 | AtREP14 | AtREP14 |
| 21_n | AtREP15 | AtREP4, 15 | AtREP15 |
| 22_v | AtREP21 | AtREP21 | AtREP21 |
| 14_o,i,l,m | AtREP3, 20 | HelitronY1E, AtREP1, 2, 2A, 3, 5, 11, 11A | AtREP3 |
| 17_p | AtREP6, 7, 8, 9, 13 | AtREP6, 7, 8, 9 | AtREP6 |
| 23_w | AtREPX1 | AtREPX1 | AtREPX1 |

Table 4: New helitron nomenclature. The first column corresponds to the new set of pairs selected through optimization. The second and third columns correspond to the former helitron name applied to these extremities [Kapitonov 2001]. The last column corresponds to the new helitron family name.

## DISCUSSION

### *Characterization of chimeric helitrons*

Many occurrences of truncated helitrons containing only one helitronic terminus were observed in the Arabidopsis genome (Table 2 and 3), thus suggesting that they were subject to incomplete excision. On the other hand, a significant number of helitrons showing a combination of helitronic termini was also observed (Figure 5), including helitrons with termini corresponding to two distinct families and/or multiple combinations of termini for unique sequences (Figure 7). Lastly, results showed that distinct helitron sequences may be bound by the same 5' and 3' terminal structures (Figure 1). It is therefore extremely difficult to propose a uniform classification of helitrons taking into account both internal sequences and the dynamics of 5' and 3' termini. It is probable, however, that this combinatorial helitron structure and its variability represent important biological properties. The insertion of truncated helitrons in the vicinity of other helitrons may be a source of structural variability, which may be ascribed to the functioning of transposition proteins, which could use a terminus from a truncated helitron and a terminus from another complete or truncated helitron [Mendiola 1994, Lai 2005]. Figures 6 and 7 suggest that the use of termini combinations is possible, although some combinations are used and combined preferentially, thus giving rise to groups that occur much more frequently than others. Truncated helitrons may therefore be an important vector of the modularity of the internal helitron sequence and/or of the creation of chimerical helitrons (Figure 1 and 9). Moreover, as

shown in Figure 9 and as observed in maize [Lai 2005], the variability and combination of sequences involve fragments of genomic DNA that are mobilized at the same time as helitrons. Therefore, in the context of such variability of internal sequences, it was noteworthy that a terminus-based analysis and classification yielded a well-structured distribution of helitron copies (Figure 7).

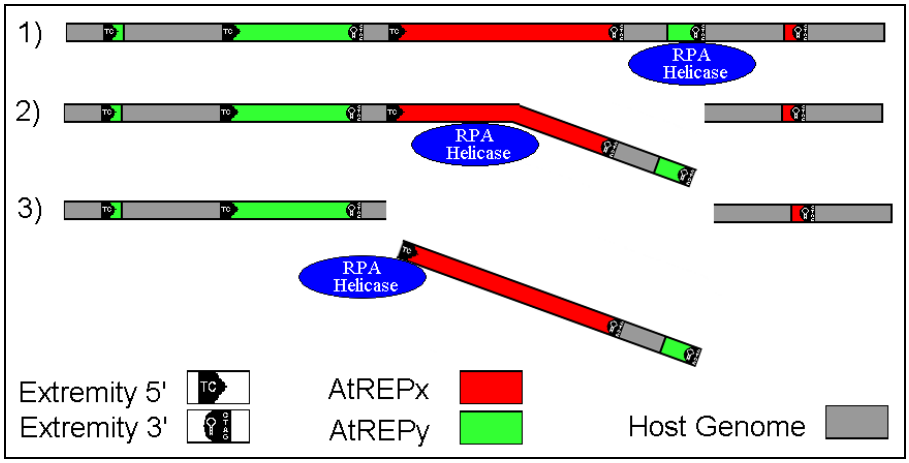


Figure 10: Hypothetical scheme of the molecular mechanisms involved in the creation of helitronic chimera (adapted from Feschotte et al. [Feschotte 2001] and Gutierrez et al. [Guiterrez 1999]). (1) A complete helitron is situated near several truncated helitrons. Transposition proteins recognize one of the 3' termini of truncated helitrons. (2) Transposition proteins cut the 3' terminus from a truncated helitron, and continue to mobilize the sequence from the 3' end towards the 5' end [Feschotte 2001]. (3) Transposition proteins recognize the 5' terminus of the complete AtREPx helitron, resulting in the transposition of a chimerical helitron.

*Relationships between helitron families and autonomous helitrons*

If the helicase-RPA transposition complex recognized a non-specific pattern in all kinds of helitrons, there would be no correlation between the number of autonomous helitrons in a given family and the amplification of this family. The comparison of internal sequences did not

show any strong correlation between the characteristics of autonomous helitrons and those of non-autonomous helitrons [Kapitonov 2001]. In contrast, the terminus-based analysis in this study highlighted significant relationships between certain autonomous helitrons and non-autonomous helitrons, which could therefore be classified in common families (Table 4). Moreover, the observed correlation between the presence of autonomous helitrons and the degree of amplification of non-autonomous helitrons belonging to the same terminus-based family strongly suggested that the RPA-helicase complex preferentially recognized helitron termini similar to those of the autonomous helitron. Some families, however, such as the AtREP3 family, consist exclusively of non-autonomous helitrons. Most of these families, except the new AtREPX1 family, have one extremity in common with one of the autonomous families. For example, the new AtREP3 non-autonomous family (combination 5' 14_o,i,l,m 3') shares the o, i, l, m 3' extremity with the new autonomous Helitron 5 family (5' 5,6_o,i,l,m 3').

Previous studies on transposon IS91, which uses a rolling-circle replication mechanism and a helicase-like transposase [Bernales 1999, del Pilar 2001], have shown that only one extremity consisting of a subterminal hairpin is necessary and sufficient for rolling-circle transposition [Mendiola 1994]. If this applied to *Arabidopsis thaliana*, the presence of a common 3' extremity may explain the amplification of non-autonomous helitrons of AtREP10, AtREP15, and AtREP3 families (Table 4) by autonomous helitrons from other families. Alternatively, the amplification of non-autonomous helitron families may have been carried out by ancient autonomous helitrons that have strongly degenerated and can no longer be detected by ORF identification and sequence analysis.

## CONCLUSION

This paper has demonstrated the significance of termini-based modeling of helitron transposable elements. This strategy provided an accurate genome-wide identification of all known sequences and resulted in the discovery of new helitron copies. Moreover, the terminus-based analysis revealed the presence of multiple termini in a significant number of autonomous and non-autonomous helitrons, thus emphasizing a novel aspect of helitron dynamics in the *Arabidopsis thaliana* genome. Finally, it revealed a highly-structured clustering of all helitron sequences that could be used for a simple and systematic classification of helitron sequences. This clustering was found to be coherent with the hypothesis that helitron transposition proteins of a given family preferentially recognize the termini of this family.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Bennetzen J.L. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Genet. & Dev, 15:621–627.

Bernales I, Mendiola M.V and De la Cruz F. 1999. Intramolecular transposition of insertion sequence IS91 results in second-site simple insertions. Mol. Microbio. 33:223-234.

Bourgeois and J.-C. Lasalle. 1971. An extension of the Munkres algorithm for the assignment problem to rectangular matrices. Communications of the ACM, 142302-806.

Brunner S, Pea G, Rafalski A. 2005. Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. Plant J. 43:799-810.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268**,** 78-94.

Cerutti H, Casas-Mollano JA. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet. 50:81-99.

Cormen T.H., Leiserson C.E., Rivest R.L., and Stein C. 2001. Introduction to Algorithms. MIT Press and McGraw-Hill, ISBN 0-262-03293-7. Section 35.3 pp.1033-1038.

Craig N.L., Gragie R., Gellert M. and Lambowitz A.M. 2002. Mobile DNA II Second Edition. ASM Press.

del Pilar Garcillan-Barcia M, Bernales I, Mendiola M.V and de la Cruz F. 2001. Single-stranded DNA intermediates in IS91 rolling-circle transposition. Molecular Microbiology. 39:494-501.

Dong S., Searls DB. 1994. Gene structure prediction by linguistic methods. Genomics. 23:540-51.

Eckardt NA. 2003. A new twist on transposons: the maize genome harbors helitron insertion. Plant Cell. 15:293-5.

Feschotte C., Mouches C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol. 17:730-7.

Feschotte C. and Wessler WR. 2001. Treasure in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc .Natl. Acad. Sci. USA. 98:8923-8924.

Feschotte C, Jiang N and Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3:329-41.

Guiterrez C. 1999. Geminivirus DNA replication. Cellular and Molecular Life Sciences. 56:313-329.

Iftode C., Daniel Y., Borowiec J.A. 1999. Replication Protein A (RPA): The eukaryotic SSB. Critical Reviews in Biochemistry and Molecular Biology. 34:140-180.

Jiang N., Bao Z., Zhang X., Eddy S.R., Wessler S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 431:569-73.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110:462-467.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA. 98:8714-9.

Kidwell, M.G. and Lisch, D.R. 2001. Perspective: transposable elements and host genome evolution. Trends Ecol. Evol. 15:95-99.

Lai J., Li Y., Messing J., Dooner H.K. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. PNAS 102 :9068-9073.

Laufs J., et al. (4 co-authors). 1995. Identification of the nicking tyrosine of *geminivirus* Rep protein. a review. Biochimie. 77:765-773.

Mendiola M.V, Bernales I. & De la Cruz F. 1994. Differential roles of the transposon termini in IS91 transposition. P.N.A.S. 91:1922-1926.

Morgante M, Brunner S, Pea S, Fengler K, Zuccolo A and Rafalski A. 2005 Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics 37:997-1002.

Munkres J. 1957. Algorithms for the Assignment and Transportation Problems. Journal of the Society of Industrial and Applied Mathematics. 5:32-38.

Nicolas, J., Durand, P., Ranchy, G., Tempel, S. and Valin, A.S. 2005. Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in genomes. Bioinformatics. 21:4408-4410.

Pritham E.J, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. P.N.A.S. 104:1895-1900.

Saitou N, Nei M. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406-25.

Searls DB. 2002. The language of genes. Nature. 420:211-7.

Tempel S, Giraud M, Lavenier D, Lerman I.S, Valin A.S, Couee I, El Amrani A and Nicolas J. 2006. Domain Organization within repeated DNA sequences: Application to the study of a family of transposable elements. Bioinformatics. 22:1948-1954.

Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-80.

Vanitharani R, Chellapan P, and Fauquet C. 2005. Geminivirus and RNA silencing. Trends in Plant Science. Review 10:144-151.