5

Computational Analysis of Transposable Element Sequences

I. King Jordan and Nathan J. Bowen

Summary

This chapter provides a simple guide for the computational analysis of transposable element (TE) sequences. Web links are provided for a number of sequence analysis applications, and their potential use in the analysis of TE sequences is briefly described. The level of detail provided is intended to be sufficient for a naive user to begin to analyze TE sequences *in silico*. The emphasis is placed on the identification, retrieval and manipulation of TE sequences. Information is also provided on the evolutionary study of TE sequences including the use phylogenetics programs.

Key Words: Transposable elements; computational analysis; sequence comparison; repeat masker; phylogenetics; alignment.

1. Introduction

The purpose of this chapter is to provide the reader with a simple heuristic guide for the computational analysis of transposable element (TE) sequences (nucleotide and/or amino acid). The current revolution in genomics has produced a wealth of sequence information. These sequence data are of particular relevance to the field of TE biology as mobile elements are ubiquitous and often constitute a substantial fraction of their hosts' genomes. Concurrent with the production of genomic sequence data has been a concerted effort to produce and disseminate the computational tools necessary to analyze and interpret these data. This work has resulted in a vast and potentially confounding array of computational tools and approaches. This chapter attempts to present a beginning framework for the appropriate selection and use of these tools. This work in no way represents a comprehensive or in-depth survey of the conceptual foundations and computational tools available for sequence analysis. Programs and analytical approaches presented here are chosen on the basis

From: *Methods in Molecular Biology, vol. 260: Mobile Genetic Elements*
Edited by: W. J. Miller and P. Capy © Humana Press Inc., Totowa, NJ

of ease of use and the authors' familiarity. An emphasis is placed on brevity, and just enough information is provided for the naive user to get started with a variety of sequence analysis techniques. Only software that is freely available, either on Web servers or as downloadable executable code, is described here. In addition, an attempt is made to provide information relevant to users working under the Windows® (PC), Macintosh®, and/or UNIX® operating systems.

Given that evolution is a unifying theme in biology and that TEs are known to have a major impact on genome evolution and organization, the emphasis here is on sequence analyses that enable detailed evolutionary inferences. However, it is hoped that the tools and approaches described here will prove to be relevant to the computational analysis of TEs in any biological context.

2. Material

URLs (Web addresses) where the programs can be found that are recommended for computational sequence analysis are given, for the most part, in **Tables 1–5** and in **Subheading 2.** The prefix http:// has been omitted from the addresses in the interest of space, but is required when typing the URL.

3. Methods

3.1. Sequence Retrieval and Manipulation

Molecular sequence data are represented as strings of characters (e.g., A, T, C, and G for nucleotides) and are stored with annotation as text files in a variety of different file formats. Some familiarity with a few of the more common sequence formats (e.g., FASTA format; *see* www.ncbi.nlm.nih.gov/BLAST/fasta.html) will be helpful to the user. Sequence analysis programs require that data be entered in specific and often different formats. Thus the use of a file-format converter will likely be inevitable in any sustained effort at sequence analysis. Format converters can input sequences (usually aligned) in a variety of formats and then output them in a different user-defined format. **Table 1** lists some programs that include file-format conversion functions.

The first step in any sequence analysis project is the retrieval of sequence data. A number of databases exist that are designed to store, organize, and disseminate sequence data (*see* **Table 2**). These include very comprehensive databases such as Genbank, more focused databases such as the TIGR microbial genome database (www.tigr.org/tdb/), and even organism-specific databases such as the yeast genome database (genome-www.stanford.edu/Saccharomyces/). Sequence retrieval from Genbank will be described here.

nphasis is placed on brevity,
ıe naive user to get started
Only software that is freely
executable code, is described
nformation relevant to users
®, and/or UNIX® operating

logy and that TEs are known
l organization, the emphasis
volutionary inferences. How-
scribed here will prove to be
ıny biological context.

can be found that are rec-
are given, for the most part,
ix http:// has been omitted
ıt is required when typing

gs of characters (e.g., A, T, C,
tion as text files in a variety
a few of the more common
w.ncbi.nlm.nih.gov/BLAST/
ıalysis programs require that
mats. Thus the use of a file-
sustained effort at sequence
usually aligned) in a variety
ser-defined format. **Table 1**
ersion functions.
is the retrieval of sequence
ıned to store, organize, and
e include very comprehen-
latabases such as the TIGR
lb/), and even organism-
database (genome-www.
val from Genbank will be

Table 1
Programs That Perform Sequence File-Format Conversion

Program	Format		URL	Operating system		
	Executable	Server	(All addresses require http:// prefix)	Windows	Macintosh	UNIX
Seqpup	X		iubio.bio.indiana.edu/soft/molbio/java/apps/seqpup/	X	X	X
DAMBE	X		web.hku.hk/~xxia/software/software.htm	X		
ClustalX	X		www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html	X	X	X
GDE	X		ftp.bio.indiana.edu/soft/molbio/unix/GDE/			X
ReadSeq		X	dot.imgen.bcm.tmc.edu/seq-util/readseq.html	X	X	X

Table 2
Molecular Sequence Databases

Database	URL (All addresses require http:// prefix)
Genbank (NCBI)	www.ncbi.nlm.nih.gov/
EMBL Nucleotide sequence database (EBI)	www.ebi.ac.uk/embl/index.html
DDBJ (DNA databank of Japan)	www.ddbj.nig.ac.jp/
SwissPROT (SIB and EBI)	www.ebi.ac.uk/swissprot/

Genbank's Entrez search and retrieval system (www.ncbi.nlm.nih.gov/Entrez/) can be used to do a variety of string searches to identify transposable element sequences of interest. For example, entering the boolean search command 'gypsy AND transposable element' in the Entrez nucleotide search field will retrieve a number of *gypsy*-like retrotransposon sequences. In addition to such string searches, the user will probably want to perform a sequence-similarity search to locate sequences that show some similarity (and thus putative relatedness) to their element of interest. This can be done using any number of different BLAST *(1,2)* searches (www.ncbi.nlm.nih.gov/BLAST/). To perform a search using BLAST, the user selects a query sequence (element of interest) to search the database of choice. The BLAST program then retrieves all sequences in the chosen database that have a similarity score at or above a user-defined value. The use of amino-acid query sequences usually results in more sensitive searches than those conducted with nucleotide query sequences, and it can effectively retrieve distantly related sequences. Another way to increase the sensitivity of a sequence search is to incorporate the site-specific information embedded in multiple alignments of related sequences. This can be accomplished using PSI-BLAST *(2)*. PSI-BLAST generates a multiple sequence alignment based on an initial BLAST search. The site-specific variation derived from this alignment is then used to iteratively re-search the database for more distantly related sequences. At each iteration the user has the option of choosing which sequences to include in the next multiple sequence alignment. The process is repeated until it converges and no new sequences are retrieved. The use of PSI-BLAST has the advantage that it can retrieve very distantly related sequences that may not be detected with a standard BLAST search. However, the increased sensitivity of PSI-BLAST can also result in more false positives, and this approach necessarily involves more input and consideration from the user.

A very powerful and useful tool designed explicitly for the identification of TEs and other repetitive sequences is the Repeat Masker program

URL
addresses require http:// prefix)

w.ncbi.nlm.nih.gov/
w.ebi.ac.uk/embl/index.html
w.ddbj.nig.ac.jp/
w.ebi.ac.uk/swissprot/

m (www.ncbi.nlm.nih.gov/
ches to identify transposable
ing the boolean search com-
ntrez nucleotide search field
on sequences. In addition to
to perform a sequence-simi-
similarity (and thus putative
be done using any number of
h.gov/BLAST/). To perform
quence (element of interest)
program then retrieves all
nilarity score at or above a
sequences usually results in
nucleotide query sequences,
sequences. Another way to
incorporate the site-specific
related sequences. This can
LAST generates a multiple
arch. The site-specific varia-
eratively re-search the data-
ch iteration the user has the
n the next multiple sequence
erges and no new sequences
dvantage that it can retrieve
be detected with a standard
ity of PSI-BLAST can also
cessarily involves more input

licitly for the identification
Repeat Masker program

Table 3
Multiple Alignment Programs

Program	Format: Executable	Format: Server	URL	Operating system: Windows	Operating system: Macintosh	Operating system: UNIX
ClustalW		X	http://www.ebi.ac.uk/clustalw/	X	X	X
ClustalW	X		ftp://ftp.bio.indiana.edu/molbio/align/clustal/	X	X	X
ClustalX	X		http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html	X	X	X
SAM		X	http://www.cse.ucsc.edu/research/compbio/HMM-apps/tuneup-alignment.html	X	X	X
MultAlin		X	http://www.toulouse.inra.fr/multalin.html	X	X	X
DIALIGN	X		http://www.gsf.de/biodv/dialign.html			X

Table 4
Phylogenetic Analysis Programs

Program	URL (All addresses require http:// prefix)	Operating system		
		Windows	Macintosh	UNIX
PHYLIP	evolution.genetics.washington.edu/phylip.html	X	X	X
MEGA	www.megasoftware.net/	X		
ClustalX	www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html	X	X	X
DAMBE	web.hku.hk/~xxia/software/software.htm	X		
Treeview	taxonomy.zoology.gla.ac.uk/rod/treeview.html	X	X	
PAUP*	paup.csit.fsu.edu	X	X	X

Treeview	taxonomy.zoology.gla.ac.uk/rod/treeview.html	X	X	
PAUP*	paup.csit.fsu.edu	X	X	X

Table 5
Sequence Divergence and Polymorphism Programs

Program	URL (All addresses require http:// prefix)	Operating system		
		Windows	Macintosh	UNIX
DnaSP	www.ub.es/dnasp	X		
MEGA	www.megasoftware.net/	X		
DAMBE	web.hku.hk/~xxia/software/software.htm	X		
JaDis	biom3.univ-lyon1.fr/software/jadis.html	X	X	X
PAML	abacus.gene.ucl.ac.uk/software/paml.html	X	X	X

(repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) (Smit, A. F. A. and Green, P., 2001, unpublished). Repeat Masker takes a FASTA-formatted sequence as input and characterizes any repetitive elements in the sequence that have similarity to the elements in its database. The "masked" regions of the input sequence, i.e., those with similarity to known repetitive DNA, are annotated and returned to the user. This program works best for systems where most if not all of the repetitive DNA elements have been defined and are thus already present in the database.

Once the user has retrieved a number of TE sequences of potential interest, the next step is to align the sequences. Multiple sequence alignment involves the identification and alignment of homologous residues among a group of related sequences, and it is a prerequisite to the extraction of meaningful biological and evolutionary information from the sequences. A number of multiple sequence alignment methods and programs are available (**Table 3**). The most commonly used program is Clustal. Clustal users can choose from two different interfaces that run the same alignment algorithm: ClustalW *(3)* has a text interface, and ClustalX *(4)* has a more user-friendly Windows interface. Sequences can be input into Clustal in a number of different formats. Prior to multiple sequence alignment, the user has the option of adjusting a number of alignment parameters, including gap penalties and the protein or DNA weight matrix to be used. Once the alignment is complete, the user can re-align selected sequences or a selected residue range. These procedures are often iterated, with variations, a number of times until the best alignment is obtained.

It is important to note that multiple sequence alignment can be quite inexact, especially with distantly related sequences. The user should always visually inspect the output from any multiple alignment program. In obvious cases, the user may choose to manually adjust any misaligned region. Some biological knowledge of the sequence in question as well as familiarity with previous work done on similar or related sequences can greatly aid in manually aligning conserved motifs. A more conservative approach would entail the removal of any ambiguous or poorly aligned region from the alignment. The reliability of any subsequent analysis depends on the accuracy of the alignment. It is therefore critical to ensure that the best possible alignment is obtained, using both the parameters of the program and manual adjustment if necessary, before proceeding with analysis of the data.

3.2. Sequence Analysis

Computational analysis of TE sequences is a comparative endeavor which at its core entails the detection and study of shared patterns among groups of related elements. The patterns that are revealed through these efforts reflect the historical process of evolution. With this relationship between pattern and

process in mind, the sequence analysis section consists of four parts. The first two parts, *Phylogenetic Analysis* and *Aging of Elements*, emphasize the detection of evolutionary patterns among elements. While the final two parts, *Genome-Level Selection* and *Host-Level Selection*, describe the study of some of the forces involved in the process of element evolution.

3.2.1. Phylogenetic Analysis

The evolutionary relationships among a related group of TEs can be discerned using phylogenetic analysis. Phylogenetic analysis can also be used to identify novel families of TEs *(5)* or to uncover recombination events between individual elements *(6)*. Phylogenetic reconstruction begins with a reliably aligned set of sequences (*see* **Subheading 3.1.**). There are three general methods of phylogenetic reconstruction: distance-based, parsimony, and maximum-likelihood *(7)*. Many programs that implement one or more of these approaches are available (**Table 4** and evolution.genetics.washington.edu/phylip/software.html).

The most commonly used programs are PAUP* *(8)* and PHYLIP *(9)*. PAUP*, as implemented on the Macintosh operating system, is probably the most user-friendly phylogenetic analysis program. However, PAUP* is not freely available. PHYLIP, although not quite as user friendly as PAUP*, is also quite useful and widely employed. PHYLIP can perform all three general methods of phylogenetic reconstruction. The program requires the user to supply an input file of aligned sequences in a format specific to the program. Alignments can be converted into this PHYLIP format using a sequence converter (**Table 1**). In order to execute a program in PHYLIP, it is simplest to put the input file in the same directory as the program to be used and name that file "infile." Use of PHYLIP can be somewhat unwieldy, as it requires each step in the analysis to be performed by a different program. For example, with distance-based phylogenetic reconstruction the user must first calculate a distance matrix using either the DNADIST or PROTDIST programs, depending on the type of sequence being analyzed.

Once a distance matrix has been calculated, it can be used with one of several tree-building programs to reconstruct the phylogeny. The most common measure of support for individual branches of phylogenetic trees is the bootstrap. As with phylogenetic reconstruction, bootstrapping with PHYLIP requires the separate use of several different programs. First, randomized replicate sequence alignments are generated using SEQBOOT. Then, for each alignment a distance matrix is calculated, followed by multiple tree reconstructions using the programs described above. Finally, a consensus phylogeny is built using the CONSENSE program. The percentage of times each branch shows up in all of the phylogenies reconstructed from the bootstrapped align-

ments can be determined from the CONSENSE output; this is taken as the measure of support for that branch. Viewing and graphically manipulating the phylogenies produced by PHYLIP and other programs can be done using the Treeview *(10)* program (**Table 4**).

3.2.2. Aging of Elements

The age of elements in the genome may be of interest to the TE researcher. Long terminal repeat (LTR)-containing retrotransposons can be aged in a straightforward way by comparing the sequences of their 5' and 3' LTRs *(11,12)*. Due to the mechanism of reverse transcription, when an LTR retrotransposon inserts into the genome its LTRs are expected to be identical in sequence. Subsequent to insertion, the LTRs accumulate mutations. Thus the level of sequence divergence between 5' and 3' LTRs of an element can be used to assess approximately how much time has elapsed since it inserted in the genome. For non-LTR elements, ages can be estimated for groups or subfamilies of related elements *(13)* as opposed to the aging of individual elements possible with LTR retrotransposons. This technique also relies on the fact that elements accumulate mutations subsequent to their insertion in the genome. In the case of a related group of non-LTR elements, the sequence of a common ancestor can be estimated either by using a consensus sequence, or by using a phylogenetic approach. Once an ancestral sequence is estimated for a given group of elements, the average number of mutations that have accumulated between the ancestral sequence and each extant sequence can be determined. This average level of sequence diversity can be used to estimate the age of a group or subfamily of elements.

3.2.3. Genome-Level Selection

In addition to the patterns of evolution revealed as described above, analysis of TE sequences can also yield information on the process of element evolution. Selection on TE sequences at the level of the genome, or inter-element selection *(14)*, occurs as a result of differential reproductive success (i.e., transposition rates) among members of a TE family. This type of selection is consistent with the "selfish DNA" hypothesis *(15,16)* of TE evolution. The role of inter-element selection can be inferred by comparing paralogous copies of elements within genomes *(17)*. Comparison of protein encoding nucleotide sequences can yield evidence of inter-element selection. In order to perform such comparisons, it is necessary to align the codons of the protein-encoding nucleotide sequences. This action can be accomplished by first aligning the encoded protein sequences, and then ensuring that the gaps in the corresponding nucleotide sequence alignment match those in the encoded amino-acid sequence alignment. This procedure is implemented in the DAMBE *(18)*

tput; this is taken as the
phically manipulating the
rams can be done using

rest to the TE researcher.
posons can be aged in a
of their 5' and 3' LTRs
cription, when an LTR
expected to be identical in
ılate mutations. Thus the
of an element can be used
d since it inserted in the
ed for groups or subfami-
g of individual elements
also relies on the fact that
insertion in the genome.
ıe sequence of a common
s sequence, or by using a
is estimated for a given
ıs that have accumulated
ıence can be determined.
l to estimate the age of a

described above, analysis
rocess of element evolu-
genome, or inter-element
uctive success (i.e., trans-
type of selection is con-
TE evolution. The role of
paralogous copies of ele-
ein encoding nucleotide
tion. In order to perform
s of the protein-encoding
hed by first aligning the
e gaps in the correspond-
the encoded amino-acid
ed in the DAMBE *(18)*

program (**Table 1**) for PC users or the DNA stacks *(19)* program (biology.fullerton.edu/deernisse/dnastacks.html) for Macintosh users.

Once the codons are properly aligned, comparison of synonymous (ds) and nonsynonymous (dn) substitution rates will yield information on the nature of selection that has operated on element sequences within genomes. A number of programs are available that can calculate ds and dn (**Table 5**) as well as other measures of nucleotide variation. Some of these programs also include more sophisticated tests that may reveal subtle effects of selection not detected by a simple ds versus dn comparison. For example, comparison of ds and dn can be done for individual branches of a phylogenetic tree to evaluate different historical episodes of selection *(20,21)*.

3.2.4. Host-Level Selection

The detection of host-level selection (i.e., between organisms) on element sequences is not as straightforward as the detection of inter-element selection and as yet is less common. The availability of identically located (orthologous) element sequences from the genomes of different but closely related species is critical to this endeavor. Unless the elements have a site-specific insertion mechanism, the presence of orthologous element sequences in related genomes indicates that these elements inserted prior to the evolutionary divergence of the genomes (species). Thus any selection acting on these sequences necessarily occurred after transposition (insertion), or in other words, during the process of species divergence *(22)*. Orthologous element sequences can be compared in the same way as described for paralogous copies above. For example, ds and dn comparisons can be made to assess whether orthologous element sequences are being conserved between species due to selection acting at the host level. In addition, nonencoding orthologous element sequences can be compared to determine if they may be conserved between species and thus potentially play some regulatory role for the host species. Selection of TE sequences at the host level is not consistent with the "selfish DNA" hypothesis of TE evolution and indicates that the element sequences in question are performing an essential function for their host species *(23)*.

References

1. Altschul, S. F., Gish, W., Miller, W., Mayers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
3. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.

4. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882.
5. Bowen, N. J. and McDonald, J. F. (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.* **9,** 924–935.
6. Jordan, I. K. and McDonald, J. F. (1999) Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the *Saccharomyces cerevisiae* genome. *Mol. Biol. Evol.* **16,** 419–422.
7. Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. Phylogenetic inference. In *Molecular Systematics* (Hillis, D. M., Moritz, C., and Mable, B. K., eds.), Sinauer Associates, Sunderland, MA, 1996, pp. 407–514.
8. Swofford, D. L. (ed.) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates, Sunderland, MA, 1998.
9. Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266,** 418–427.
10. Page, R. D. M. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12,** 357–358.
11. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20,** 43–45.
12. Promislow, D. E., Jordan, I. K., and McDonald, J. F. (1999) Genomic demography: a life-history analysis of transposable element evolution. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 1555–1560.
13. Kapitonov, V. and Jurka, J. (1996) The age of Alu subfamilies. *J. Mol. Evol.* **42,** 59–65.
14. McDonald, J. F., Matyunina, L. V., Wilson, S., Jordan, I. K., Bowen, N. J., and Miller, W. J. (1997) LTR retrotransposons and the evolution of eukaryotic enhancers. *Genetica* **100,** 3–13.
15. Doolittle, W. F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284,** 601–603.
16. Orgel, L. E. and Crick, F. H. (1980) Selfish DNA: the ultimate parasite. *Nature* **284,** 604–604.
17. Jordan, I. K. and McDonald, J. F. (1999) Comparative genomics and evolutionary dynamics of Saccharomyces cerevisiae Ty elements. *Genetica* **107,** 3–13.
18. Xia, X. (ed.) *Data Analysis in Molecular Biology and Evolution.* Kluwer Academic Publishers, Norwell, MA, 2000.
19. Eernisse, D. J. (1992) DNA Translator and Aligner: HyperCard utilities to aid phylogenetic analysis of molecules. *Comput. Appl. Biosci.* **8,** 177–184.
20. Messier, W. and Stewart, C. B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* **385,** 151–154.

21. Zhang, J., Rosenberg, H. F., and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95,** 3708–3713.
22. Witherspoon, D. J., Doak, T. G., Williams, K. R., Seegmiller, A., Seger, J., and Herrick, G. (1997) Selection on the protein-coding genes of the TBE1 family of transposable elements in the ciliates *Oxytricha fallax* and *O. trifallax*. *Mol. Biol. Evol.* **14,** 696–706.
23. Miller, W. J., McDonald J. F., and Pinsker, W. (1997) Molecular domestication of mobile elements. *Genetica* **100,** 261–270.