*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

*Sequence analysis*

# CombTEs: a result from comparing pHMMs and the library-based method to search for LTR retrotransposons

Carlos N. Fischer

Department of Statistics, Applied Maths, and Computer Science, UNESP - São Paulo State University, 13506-900 Rio Claro, SP, Brazil.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

**Abstract**

**Motivation:** A method commonly used to identify Transposable Elements (TEs) is the library-based one that uses a set of reference sequences to search for similar repeats in a sequence. The main drawback of this method is that it is able to find only repeat copies of previously identified TEs. Profile hidden Markov models (pHMMs) can be an interesting alternative, particularly when repeats in the query sequence present low or even no similarity with the used reference sequences.

**Results:** This work presents a comparison between results produced by these two methods to search for LTR retrotransposons, showing that the pHMMs can predict more annotated TEs (reaching 37%) than the library-based method. Each method was able to predict annotated TEs missed by the other one: from the total of correctly predicted TEs, about 28% were predicted exclusively by the pHMMs and 1.3% were found only by the similarity method. Considering this, this work also presents a tool that combines results produced by the library-based and pHMMs methods on the search for LTR retrotransposons. Combining results produced by different approaches, besides increasing the number of correct predictions, is a way to improve the reliability of the final findings.

**Availability:** Data and software are available at https//: .

**Contact:** carlos.fischer@unesp.br

**Supplementary information:** Available at *Bioinformatics* online.

## Introduction

The identification and classification of Transposable Elements (TEs) is an important step in genome annotation. The characterization of TEs and their influence on a genome have been treated at length in several works (Kidwell and Lisch, 2001; Solyom and Kazazian, 2012). LTR retrotransposons (LTR-RT) are one of the classes of TEs - Bel-PAO, Copia, and Gypsy superfamilies are the focus in this work. Several methods aimed to identify and/or classify TEs have been cited in the literature (Bergman and Quesneville, 2007; Lerat, 2010); library-based approach is the most commonly used one. With respect to TEs, this method consists in using a set of reference sequences of TEs to search for similar repeats in a query sequence. Its main drawback is that it can find only repeat copies of previously identified TEs. Despite of this drawback, the usage of this method is often the first step in TE identification. An alternative is the profile hidden Markov models (pHMMs) (Eddy, 2009), probabilistic models that can be used to search for members of a particular class of sequences. In the case of TEs, the pHMM of a specific superfamily is used to search for repetitions of that superfamily into a genome, including distant homologue copies (Eddy, 2009). It means that pHMMs can be particularly interesting when repeats in the query sequence present low or even no similarity with the reference sequences.

This work has two aims. The first one is to present a comparison between results from running the library-based method and pHMMs on a genome, showing that the latter can perform better than the first one in searches for LTR-RT. The second objective is to present a tool that combines results produced by these two methods (and also others). Combining results produced by different approaches, in addition to increasing the number of correct predictions, is a way to further improve the reliability of the TEs identified in a query sequence.

## Results

The tests with the two methods were performed using RepeatMasker (Smit et al., 2010), an efficient and widely used tool that applies the library-based approach, and HMMER (Eddy, 2009), a program that uses pHMMs. Both programs were run on the *Drosophila melanogaster* genome, downloaded from FlyBase (http://flybase.org/), together with its annotation, used as the official one (describing 4162 sequences of Bel-PAO, Copia, and Gypsy). The reference sequences of LTR-RT were obtained from Repbase (Jurka et al., 2005). Important to observe that sequences of any *Drosophila* were excluded from the sets used to build the libraries for RepeatMasker and train the pHMMs - in the tests were used three different pHMMs for each superfamily (Fischer et al., 2018). All this material provided the necessary conditions and results to allow the evaluation of the tested approaches. Details about all this, can be found in Section S1 of the Supplementary Material (SM).

RepeatMasker uses the Smith-Waterman (SW) score to evaluate the match between a sequence from its library and the query sequence; for HMMER, an E-value is reported as the statistical significance of a match (the lower the E-value, the more significant the hit). The initial results were obtained by running the tools with their default parameter settings (225 is the default value for RepeatMasker).

Table 1 presents the numbers of false positive (FP) predictions and missed annotations (FN) produced in the tests - the initial results are labeled “No filter” in the “Filter of” columns; the other numbers were obtained by filtering the SW scores or E-values of the initial results. The table shows that, from a total of 4162 annotations, RepeatMasker missed 427 (24.5%) annotations more than HMMER (“No filter” line) and that both tools predicted very high numbers of FP. However, the table also shows that, by applying filters on the initial results, the numbers of FP were drastically reduced (but increased the number of missed annotations - FN columns). For values of filters which maintain very close the amounts of FP for both methods (the last four lines of the table), the pHMMs always produced lower numbers of missed annotations: between 552 and 291, depending on the used filters. Also, from the total of TEs correctly predicted by both tools, between 28% and 22% were predicted exclusively by the pHMMs and between 1.3% and 4.9% were found only by the similarity method.

Thus, w.r.t. the first aim of this work, the conclusion is that the pHMMs presented better performance (37.3% to 22.3% more correct predictions) than the library-based method, considering the used tools (details can be found in Section S2 of SM).

**Table 1.** Numbers of False Positive (FP) predictions, missed annotations (FN) and combined predictions.

| pHMMs (tamanho 9) | | | Library-based method | | | Three | Two | One |
|---|---|---|---|---|---|---|---|---|
| Filter of (for E-values): | FP [common FP] | FN | Filter of (for SW scores): | FP [common FP] | FN | Total of combined predictions [% of correct ones] | | |
| No filter | 17,783 | 1,744 | No filter | 12,948 | 2,171 | 2,048 [42.5%] | 1,309 [33.6%] | 24,738 [6.5%] |
| 1e-05 | 1,010 [51.7%] | 2,131 | 393 | 1,058 [51.9%] | 2,683 | 1,829 [71.3%] | 474 [60.3%] | 1,025[26.4%] |
| 1e-10 | 867 [59.4%] | 2,239 | 405 | 871 [61.6%] | 2,704 | 1,762 [71.9%] | 386 [62.2%] | 801 [34.1%] |
| 1e-20 | 684 [68.8%] | 2,426 | 440 | 701 [70.6%] | 2,783 | 1,513 [71.9%] | 351 [69.2%] | 635 [47.4%] |
| 1e-30 | 595 [73.9%] | 2,564 | 480 | 592 [77.7%] | 2,855 | 1,313 [73.4%] | 364 [67.6%] | 571 [56.7%] |

Regarding the FP predictions, the corresponding columns of Table 1 show the percentages (in brackets) of FP of a method that are also FP of the other one for the tested filters. This may indicate that many FP (at least 50%) could be, in fact, correct predictions that would be missing (or misannotated) in the official annotation.

Based on this, the second aim of this work is to present a Perl script that combines predictions produced by HMMER and RepeatMasker; it also allows to consider results from other tools. Its output files are named corresponding to the number of tools used to produce each combined prediction. This script was tested considering also results from RPS-Blast (Camacho, 2017) (this program also reports an E-value for each found domain); these new results are also shown in Table 1 (columns named Three, Two, and One) - each line presents the total number of resulting combined predictions and the percentage of correct ones according to the official annotation (the same filter used for HMMER was considered for RPS-Blast).

Regarding the combined predictions produced by all the three tools, “Three” column shows that more than 71% of them are correct ones for all used filter associations; the rest would be initially considered as FP. However, observing the E-values and SW scores of these FP, a very high percentage of these FP present at least one of these metrics as being good one (i.e., low E-value and/or high SW score): for the first filter association, 525 (28.7% of 1829) combinations would be FP but, for the ones that consider a prediction as a correct one when its E-value for HMMER or RPS-Blast is at maximum 1e-20 and/or its SW score are at minimum 500, 98.7% of those 525 FP would be correct ones; for 1e-30 and 600, respectively, that percentage reaches 96% - equivalent analysis can be conducted for “Two” column (any two of those tools) (see more

in Section S3 of SM and about resulting lists of TEs possibly missing in the official annotation or that present annotation problems). These high percentages show that the implemented script can be an interesting auxiliary tool when using two or more approaches to search for LTR-RT, making the succeeding analysis of the combined results easier and faster.

In the implemented script, the user can set filters and other parameters to select the initial predictions for producing the resulting combinations. Also, in the performed tests, the script's run on the biggest chromosome took less than 3 seconds in a conventional computer; this means that, after running just once the tools to search for TEs, the user can do very fast and easily several tests with different filter combinations.

The main script is made available together with others aimed at extracting and formating the raw results from HMMER, RepeatMasker, and RPS-Blast for the process of prediction combination. The output files are presented in a specific format (very easy to read and analyze) and also in tabular format. Additionally, it creates a file with the TE candidates from each used tool (details in Section S4 of SM).

***Funding***: This work has been supported by the São Paulo Research Foundation-FAPESP (Grant No. 2012/24774-2).

## References

Bergman, C.M., and Quesneville, H. 2007. Discovering and detecting transposable elements in genome sequences. Brief. Bioinform. 8, 382-392.

Camacho, C., BLAST+ Release Notes. Updated 2017 Jan 6. In: BLAST Help [Internet]. Bethesda (MD): Nat.Center for Biotechnology Information (US).

Eddy, S.R. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23, 205-211.

Fischer, C.N.; Campos, V.A.; Barella, V.H. On the search for retrotransposons: alternative protocols to obtain sequences to learn profile hidden Markov models. J. Comput. Biol. 2018, 25. https://doi.org/10.1089/cmb.2017.0219.

Jurka, J., Kapitonov, V.V., Pavlicek, A., et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462-467.

Kidwell, M.G., and Lisch, D.R. 2001. Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution. 55, 1-24.

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. Heredity. 104, 520-533.

Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0; 2010.

Solyom, S., and Kazazian, H.H. 2012. Mobile elements in the human genome: Implications for disease. Genome Med. 4:12. https://dx.doi.org/10.1186/gm311.