# Annotating transposable elements in the genome using relational decision tree ensembles

Eduardo P Costa[1], Leander Schietgat[1], Ricardo Cerri[2], Celine Vens[1], Carlos N Fischer[3], Claudia M A Carareto[4], Jan Ramon[1], and Hendrik Blockeel[1,5]

[1] Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium
[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Campus de São Carlos, 13560-970 São Carlos, SP, Brazil
[3] Department of Statistics, Applied Mathematics, and Computer Science, UNESP São Paulo State University, Avenida 24-A, 1515, 13506-900 Rio Claro, SP, Brazil
[4] Department of Biology, UNESP São Paulo State University, Cristóvão Colombo, 2265, 15054-000 São José do Rio Preto, SP, Brazil
[5] Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

**Abstract.** Transposable elements (TEs) are DNA sequences that can change their location within the genome. They contribute to genetic diversity within and across species and their transposing mechanisms may also affect the functionality of genes. Accurate annotation of TEs is an important step towards understanding their effects on genes and their role in genome evolution. We introduce a framework for annotating TEs which is based on relational decision tree learning. It allows to naturally represent the structured data and biological processes involving TEs. Furthermore, it also allows the integration of background knowledge and benefits from the interpretability of decision trees. Preliminary experiments show that our method outperforms two state-of-the-art systems for TE annotation.

**Keywords:** relational decision trees, hidden Markov models, genome annotation, transposable elements

## 1 Introduction

Transposable elements (TEs) are DNA sequences that can change their location within the genome. This transposition process is carried out by a copy-and-paste (Class I TEs) or cut-and-paste (Class II TEs) mechanism. TEs make up a large portion of the DNA in eukaryotic organisms and contribute to genetic diversity within and across species. Furthermore, their transposing mechanisms increase the size of the genome and may affect the functionality of genes. Accurate annotation of TEs, together with the development of interpretable models explaining these annotations, is an important step towards understanding their effects on genes and their role in genome evolution [1].

Currently, the annotation of TEs involves a fair amount of manual labor. Automated methods exist that screen DNA for candidate TEs, but human annotators take over from there. In this paper, we explore how inductive logic

programming (ILP) can be used to improve the screening. The framework we propose uses existing methods to create a logic-based representation for each sequence, and then applies an ILP model. A preliminary experimental evaluation reveals that this method can substantially boost the precision of the screening process.

## 2 Background

As discussed by Bergman and Quesneville [2], there are several methods to annotate TEs in a genome, either based on homology, structural information or no prior information at all, i.e., by looking for repeats in the genome (known as *de novo* repeat discovery methods). However, all these approaches suffer from limitations, such as the assumption that TE sequences are very similar, the propagation of incorrect annotations, and specificity towards particular TE (super)families. Moreover, none of them automatically learn models from data, with the exception of Loureiro et al. [3], who demonstrate that machine learning methods can be used to boost the annotation of TEs. They assessed a set of annotation tools and learn a classifier that combines their predictions. A second classifier predicts which tool to use to determine the boundaries of a TE.

From biology, it is known that TEs consist of several subsequences, called *protein domains*, which help the TE perform its biological functions, including copying or moving the TE. Methods exist for recognizing protein domains; the state of the art uses hidden Markov models (HMMs) [4]. While TEs typically contain particular protein domains, the occurrence of some domains does not guarantee that a sequence is a TE.

In this work, we focus on predicting LTR retrotransposons, a particular type of TEs that belong to Class I. They are characterized by having long terminal repeats (LTRs) at their boundaries. These are identical sequences (up to minor variations) of a few hundred nucleotides that can easily be recognized using existing software [5]. There are different types of LTR retrotransposons, organized into superfamilies and families. One can consider two tasks: (1) identifying LTR retrotransposon sequences in the genome, and (2) predicting their (super)family.

## 3 Method

Given the biological background knowledge mentioned above, it makes sense not to try to learn a model that identifies TEs from the nucleotide sequence alone, but to use the tools that already exist to extract relevant information from sequences. Therefore we propose the following three-step framework for identifying TEs.

1. The genome is screened for potential LTR retrotransposons. To that aim, we use the tool LTR_FINDER [5], which scans a DNA sequence to search for matching string pairs (the LTRs), and then filters the list by checking user defined length restrictions. Each remaining candidate, i.e., the region

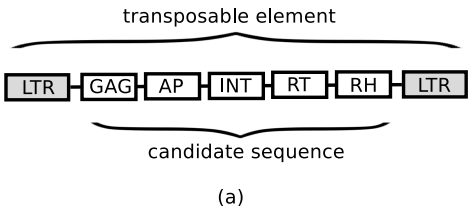


**Fig. 1.** (a) Illustration of the typical structure of a TE from the Copia superfamily, delimited by LTRs and annotated with protein domains [7]. (b) An example of an interpretation, which consists of protein domain predictions. For each domain prediction, we have the Copia candidate ID, the domain, the start and end positions of the domain prediction in the sequence, and the e-value for the HMM prediction. Note that domains may have subtypes. For example, RT_LTR is a subtype of domain RT.

bounded by the LTR pairs, receives a score, depending on how many of a predefined set of structural elements are found in there. The output of this first step is a list of candidate LTR retrotransposons, to be further filtered.

2. Every candidate TE sequence, obtained in the previous step, is screened for the occurrence of protein domains that are known to occur in LTR retrotransposons. Domains are recognized using a profile hidden Markov model (HMM) [4] trained on a multiple sequence alignment corresponding to that subdomain, from the Conserved Domain Database (CDD) [6].
3. Each candidate sequence is represented in a first order logic format, by simply listing all its predicted protein domains, and the location in the sequence where that domain was found. Fig. 1 illustrates this representation. For a given sequence, this representation is fed into an ILP model, together with biological background knowledge. The model predicts for every LTR retrotransposon superfamily the probability that the sequence belongs to that family.

In the above description, every element of this prediction framework is determined, except for the ILP model, which is to be learned from data. In our approach, the learning process is as follows. For each LTR retrotransposon superfamily, a separate model is learned that maps a sequence, represented as above, to the probability that the sequence belongs to that superfamily. This model is built using the FORF approach (first-order random forests) [8], as implemented in the relational data mining system ACE[1]. The language bias that is used allows for the following types of tests in the nodes of the tree: (1) the occurrence of a particular protein domain, (2) the occurrence of a particular protein domain with a certain length limit (the same domain can be predicted with different lengths), (3) the occurrence of a particular protein domain before another domain, and (4) the number of occurrences of a particular protein domain. As domains may have subtypes (see Fig. 1), we give the hierarchical "is a subtype" relationship as background knowledge.

[1] `http://dtai.cs.kuleuven.be/ACE/`

## 4 Experiments

As a first experiment, we evaluate the predictive performance of our framework on the genome of *D. melanogaster*[2] and compare our results for the Copia superfamily with those of two state-of-the-art methods for TE annotation: RepeatMasker and LtrDigest.

**State of the art.** RepeatMasker [9] is a tool used to find repeats in query sequences according to their similarity with sequences from a library. This program has been widely used for screening the genome for candidate TEs. LtrDigest [10] is more similar to our method. It starts from a generated list of LTR retrotransposons [11] and further annotates the sequences with protein domains (also using profile HMMs) and other structural regions. Using a fixed set of rules, a DNA strand is assigned to the most confident predictions.

**Methodology and parameter settings.** In order to obtain the initial set of candidate sequences, we used standard parameters for Ltr_Finder, but chose the lowest possible score threshold (to have as many candidates as possible) and set the minimum length between LTRs to 100 nucleotides. To annotate protein domains, we used HMMER3 [12] with default parameters. To construct a training set for FORF, we extracted 4710 positive sequences (annotated with Copia) from NCBI[3] and added the same amount of negative sequences that were similar to at least one positive sequence, but not annotated with Copia. The relational trees were built with default parameters, except for the minimum number of examples in a leaf, which was set to 5. RepeatMasker and LtrDigest were run with their standard parameter settings. For LtrDigest, we only retained predictions with an assigned DNA strand.

We report the performance of the different methods with precision-recall (PR) curves [13]. We consider a prediction correct if its boundaries deviate no more than 500 nucleotides from the boundaries of the corresponding annotation in the genome. The motivation for using PR curves is that, as only a small fraction of the genome contains Copia TE sequences, we are more interested in recognizing the positive instances than in correctly predicting the negatives.

**Results.** The PR-curves are shown in Fig. 2. A first observation is that the curve has a maximal recall of 0.63. This is due to the candidate set returned by Ltr_Finder: after removing duplicate sequences (Ltr_Finder may return sequences that only differ in the length of their LTR), it returns 4652 sequences, containing only 32 of the 51 known Copia TE sequences in *D. melanogaster*. Interestingly, 8 of the 19 missed sequences do not have any protein domain and are much shorter than a typical Copia sequence, which makes it difficult for many methods to detect them.

Second, the FORF curve is remarkable: the different points merely form a vertical line. This means that, at the highest possible threshold, the method is able to detect (nearly) all true positives. Allowing more predictions by lowering

[2] We use version 48 of the annotated genome from Flybase (http://flybase.org/), as the official annotation, which was made publicly available in November 2012.

[3] National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov/`.

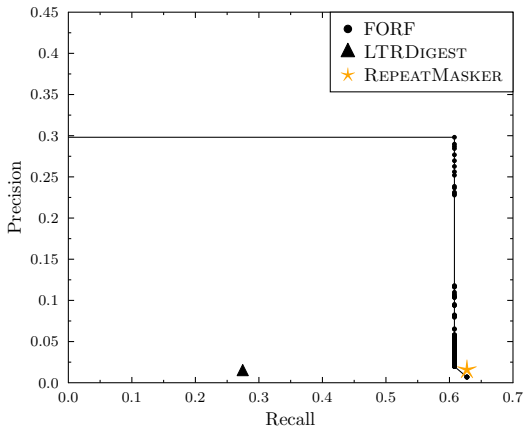


**Fig. 2.** Precision-recall curves of the different methods.

the threshold only adds false positives. This suggests that a very high cut-off can be used. We therefore had a closer look at the top predictions of the random forest (73 predictions, after removing the true positives). This resulted in a set of 6 Copia candidate sequences that are not in the official *D. melanogaster* annotation, but for which Copia evidence can be found using a BLAST search against the Nucleotide Collection (nt) of NCBI. This shows that our method is able to make new suggestions for annotations to be included in the official list.

Third, as RepeatMasker (2064 predictions) and LtrDigest (1051 predictions) only output positive predictions with 100% confidence rather than probabilities, they correspond to a single point in PR space. As the point of LtrDigest is below the curve of FORF, the latter is able to obtain a higher recall (precision) than LtrDigest for the same precision (recall). RepeatMasker yields a precision (0.02) that is slightly higher than the precision of FORF (0.01) at a recall of 0.63, but for a slight decrease of 0.02 in recall, FORF's precision rises to 0.30.

## 5 Conclusions and further work

In this paper we have proposed a framework based on relational learning to annotate TEs in a genome. We evaluated our approach for the Copia superfamily in *D. melanogaster* and found a much better predictive performance compared to the state-of-the-art methods.

As future work, we plan to explore alternatives to Ltr_Finder to select candidate sequences, in order to identify more Copia TEs contained in the official *D. melanogaster* annotation. We also plan to evaluate our method on other genomes,

which will allow us to analyze how robust our trained models are w.r.t. different organisms. Finally, we will explore hierarchical classification methods that can exploit the underlying structure of the TE classification scheme.

## Acknowledgments

We thank the Explorative Scientific Co-operation Programme between KU Leuven and São Paulo State University (UNESP). E.P.C. was supported by project G.0413.09 "Learning from data originating from evolution", funded by FWO-Vlaanderen. L.S. is supported by ERC Starting Grant 240186 "MiGraNT: Mining Graphs and Networks: a Theory-based approach" and the Research Fund KU Leuven. R.C. is funded by São Paulo Research Foundation (FAPESP - Brazil), process 2009/17401-2. C.V. is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO-Vlaanderen). C.M.A.C. is a researcher from the National Council for Scientific and Technological Development (CNPq-Brazil).

## References

1. Wheeler, T.J., et al.: Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic acids research **41**(D1) (2013) D70–D82
2. Bergman, C.M., Quesneville, H.: Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics **8**(6) (2007) 382–392
3. Loureiro, T., Camacho, R., Vieira, J., Fonseca, N.: Boosting the detection of transposable elements using machine learning. Advances in Intelligent Systems and Computing **222** (2013) 85–91
4. Eddy, S.R.: Profile hidden Markov models. Bioinformatics **14**(9) (1998) 755–763
5. Xu, Z., Wang, H.: LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research **35**(suppl 2) (2007) W265–W268
6. Marchler-Bauer, A., et al.: CDD: a conserved domain database for protein classification. Nucleic Acids Research **33**(suppl 1) (2005) D192–D196
7. Wicker, T., Sabot, F., Hua-Van, A., et al.: A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics **8** (2007) 973–982
8. Van Assche, A., Vens, C., Blockeel, H., Džeroski, S.: First order random forests: Learning relational classifiers with complex aggregates. Machine Learning **64**(1-3) (2006) 149–182
9. Smit, A., Hubley, R., Green, P.: Repeatmasker open-3.0 (2010) `www.repeatmasker.org`.
10. Steinbiss, S., Willhoeft, U., Gremme, G., Kurtz, S.: Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research **37**(21) (2009) 7002–7013
11. Ellinghaus, D., Kurtz, S., Willhoeft, U.: LTRharvest, a efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf. **9**(18) (2008)
12. Eddy, S.R.: A new generation of homology search tools based on probabilistic inference. Genome Informatics **23**(1) (2009) 205–211
13. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proc. of the 23rd Int'l Conference on Machine Learning. (2006) 233–240