BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 10 2009, pages 1329–1330
doi:10.1093/bioinformatics/btp084

Sequence analysis

TEclass—a tool for automated classification of unknown eukaryotic transposable elements

György Abrusán[1,2,*], Norbert Grundmann[2], Luc DeMester[1] and Wojciech Makalowski[2]

[1]Katholieke Universiteit Leuven, Department of Biology, Laboratory of Aquatic Ecology and Evolutionary Biology, Ch. Deberiotstraat 32, 3000 Leuven, Belgium and [2]University of Münster, Faculty of Medicine, Institute of Bioinformatics, Von-Esmarch-Str. 54, D-48149 Münster, Germany

Received on December 1, 2008; revised on January 13, 2009; accepted on February 9, 2009

Advance Access publication April 5, 2009

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The large number of sequenced genomes required the development of software that reconstructs the consensus sequences of transposons and other repetitive elements. However, the available tools usually focus on the accurate identification of raw repeats and provide no information about the taxonomic position of the reconstructed consensi. TEclass is a tool to classify unknown transposable elements into their four main functional categories, which reflect their mode of transposition: DNA transposons, long terminal repeats (LTRs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). TEclass uses machine learning support vector machine (SVM) for classification based on oligomer frequencies. It achieves 90–97% accuracy in the classification of novel DNA and LTR repeats, and 75% for LINEs and SINEs.

Availability: http://www.compgen.uni-muenster.de/teclass, stand alone program upon request.

Contact: abrusan@uni-muenster.de

1 INTRODUCTION

Transposable elements (TEs) are present in the vast majority of multi-cellular organisms. Their correct identification is a critical step in the annotation of newly sequenced genomes. However, in order to annotate repeats, their consensus sequences first have to be identified. Traditionally TE consensi were reconstructed manually, but in recent years several tools have been developed to reconstruct TEs in newly sequenced genomes, for example RepeatScout (Price *et al.*, 2005), RECON (Bao and Eddy, 2002) or RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), which together with other tools integrates them into one pipeline. The identification of repetitive sequences usually results in the raw TE consensus sequences, but does not provide information about the type or mechanism for transposition of the reconstructed repeat. So far only a few tools, for example RepeatModeler, make an attempt to classify the newly reconstructed consensi, using sequence similarity to known repeats. Similarity based classification of TEs is efficient for TEs which originate from species which have repeats closely related to other known repeats. As the price of sequencing drops, more and more species are sequenced, especially from the so far neglected parts of the phylogenetic tree. However, in the case of taxonomic groups which until now received less attention, newly identified TEs frequently show no clear similarity to known repeats, and thus their classification requires other approaches.

The problem of classification of novel repeats is largely similar to the classification/assembly problems of microbial metagenomic research—since the vast majority of microorganisms are unculturable at present, and a very large fraction of the newly sequenced microbial DNA shows no sequence similarity to sequences of known organisms. One solution for this problem is using oligomer profiles during the classification (McHardy *et al.*, 2007) and assembling the sequences with a similar profile, since the oligomer composition of many organisms is distinct. TEs were reported to have different sequence composition than genes (Andrieu *et al.*, 2004), and we have developed a simple and fast tool that uses a machine learning approach to classify unknown repetitive elements using the oligomer frequencies of the repeats. The tool (TEclass) can classify unknown TEs to their main taxonomic branches, which also reflect their mechanism of transposition: DNA transposons, Long Terminal Repeats (LTRs) and non-LTR repeats: long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).

2 METHODS

The classifiers were built using TE sequences available in RepBase (Jurka *et al.*, 2005, RepeatMasker edition), the largest database of eukaryotic repetitive sequences. Since many entries in RepBase are highly similar to each other, and sometimes represent only different evolutionary stages of the same TE lineage, for the sequences that are more than 90% similar to each other we used only the longest one during classifier building. The length of TEs varies almost two orders of magnitude, from a few hundred bases to well above 10 kb, and many repeat types have characteristic length ranges. We analyze repeats in different size categories: 0–600, 601–1800, 1801–4000 and >4000 bp and built independent classifiers for all these length classes. We use LIBSVM (Chang and Lin, 2001) as the support vector machine (SVM) engine, with a Gaussian kernel. The classification process is binary, with the following steps (Fig. 1): forward versus reverse sequence orientation > DNA versus Retrotransposon > LTRs versus non-LTRs for retroelements > LINEs versus SINEs for non-LTR repeats. The last step is performed only for repeats with lengths below 1800 bp, because we are not aware of SINEs longer than 1800 bp. Separate classifiers were built for each length class and for each classification step. In each classification step, the sequence of a TE is

[*]To whom correspondence should be addressed.

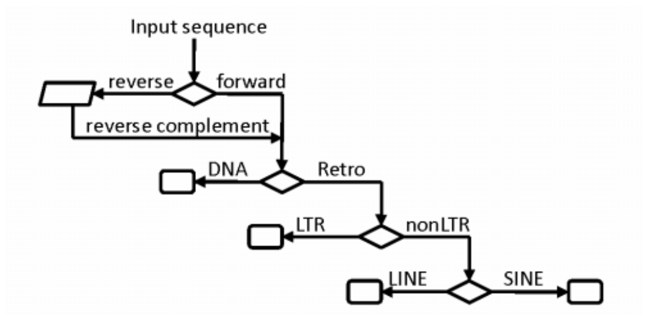

Fig. 1. Classification steps of TEclass. If the tested TE sequence is classified as reverse, it is reverse-complemented, and the subsequent steps are performed on this sequence.

represented as a vector of oligomer frequencies, which was used as the input for the SVM engine.

Two complete sets of classifiers were built using tetramers and pentamers, which are used in two separate rounds of the classification. The first step of the analysis is different from the rest both in the building of the classifiers and in the classification itself, because RepBase contains only sequences in the forward direction, but one cannot assume the same a priori about the tested TE consensi. The classifiers used in the first step were built with the RepBase repeats and their reverse complemented sequences. If an unknown repeat is classified as reversed, the further steps of the classification are performed with its reverse complemented sequence. Repeat identification is performed in two rounds. In the first, the models based on tetramer frequencies are used, and in the second, round the models based on pentamers. The result of the classification is the last step where the two rounds are in agreement, i.e. if the first classification round classifies a TE as LTR while the second as LINE it is reported as a retroelement.

3 RESULTS AND DISCUSSION

Cross validation efficiency for the different classifiers varies between 77% and 97%, with the lowest efficiency in the forward versus reverse split. We found no dramatic difference between the performance of the models based on tetra, or pentamers. In most cases, selection of a subset of the oligo-features did not result in improved classification efficiency, thus the models include the full sets of oligomers (256 tetramers and 1024 pentamers). The classification efficiency for 'unknown' repeats was determined as follows: first, the classifiers were built using the 12.11 version of RepBase (released on 14 December , 2007), and the efficiency of classification was determined for the repeats that were added to the database later, until the 13.06 release (1 August , 2008; 1604 new repeats). The performance of TEclass is different for different repeat types; >90% of the DNA transposons and LTRs were classified correctly (Table 1), while on non-LTR repeats it achieved only ~75%. The lower sensitivity of LINE/SINE classification is mainly due to the accumulation of errors during the classification process. Alone, LINEs and SINEs can be separated accurately: the cross validation efficiency is 92.4% for short (<600 bp) and 96.8% for medium length repeats (600–1800 bp). Note that this classification is not performed for repeats longer than 1800 bp, because SINEs are shorter than that. However, before this classification happens, a TE consensus sequence has to be also classified as a retroelement and subsequently as a non-LTR repeat; all these steps are error prone. We also classified the entire 13.06 release of RepBase with classifiers built using the same repeats; using these classifiers, TEclass achieved almost 100% classification efficiency (Table 1), only ~1.1% of the repeats could not be classified. This proves the high potential of the presented method and the usefulness in the annotation of new genomes forged with poorly characterized repetitive elements. The most significant shortcoming is probably that it cannot distinguish between transposable and non-TEs, thus assumes that all input sequences are TE consensi. Most tandem repeats, tRNAs and many satellites can be safely identified before building a putative TE consensus sequence, however, if very abundant, duplications of non-coding sequence, which can have essentially any sequence composition may be reported as putative TE consensi by TE reconstruction tools like RepeatScout or RECON. Separating such non-TE but repetitive sequence from TEs seem to be impossible with SVM classifiers.

Table 1. Classification efficiency of TEs

	Classifiers 2007		Classifiers 2008	
	No.	Percentage correct	No.	Percentage correct
DNA	417	90.9	2323	99.9
Retroelements	988	97.1	5646	99.8
LTR	860	94.3	4303	99.9
Non-LTR	128	75	1319	99.8
LINE	112	74.1	942	99.7
SINE	16	81.2	377	99.7
All classified	1405	91.5	7969	99.9
Indecisive	198		89	

The first test set was the sequences that were added between 14 December 2007 and 1 August 2008 to RepBase, and these classifiers were built independently using the 2007 edition of RepBase, which did not contain these sequences. We also classified all TEs in the 2008 edition of RepBase, with classifiers built with the same repeats.

Funding: Katholieke Univeristeit Leuven, Belgium (post-doctoral fellowship to G.A.); University of Münster, Germany.

Conflict of Interest: none declared.

REFERENCES

Andrieu,O. *et al.* (2004) Detection of transposable elements by their compositional bias. *BMC Bioinformatics*, **5**, 94.

Bao,Z and Eddy,S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines (version 2.86, 2008). Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Jurka,J. *et al.* (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.

McHardy,A.C. *et al.* (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.

Price,A.L. *et al.* (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21**, i351–i358.