Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 668–674

Expert Systems with Applications

www.elsevier.com/locate/eswa

Detecting LTR structures in human genomic sequences using profile hidden Markov models

Li-Ching Wu [a], Hsien-Da Huang [b], Yu-Chung Chang [c], Ying-Chun Lee [d], Jorng-Tzong Horng [a,d,*]

[a] *Institute of System Biology and Bioinformatics, National Central University, Taiwan*

[b] *Department of Biological Science and Technology, Institute of Bioinformatic National Chiao-Tung University, Taiwan*

[c] *Department of Biotechnology, Ming Chuan University, Taiwan*

[d] *Department of Computer Science and Information Engineering, National Central University, Taiwan*

Abstract

More than 45% of human genome has been annotated as transposable elements (TEs). The human genome is expanded by the mobilization of these TEs, which they may increase the plasticity and variation of the genome. Long terminal repeat (LTR) retrotransposons are important components in TEs. LTRs include regulatory sites, which the authors believe could be conserved in evolution. Therefore, these significant motifs in the sequence of LTRs are found and are used to train a Hidden Markov Model. These models are used as fingerprints to detect most of the known LTRs detected by RepeatMasker. LTR instances are classified into families using the predictive models proposed. These LTRs can support evolutionary analysis. A new method of detecting LTR is proposed. Analyzing LTR sequences reveals some specific motifs as LTR fingerprints, which can be built into HMM profiles. Experimental results reveal that the proposed experimental approach not only discovers most of the LTRs found by RepeatMasker, but also detects some novel LTRs. Moreover, the novel LTRs may be structurally incomplete or degenerate.

Keywords: Genome; Hidden Markov model; LTR; Repeats; Transposable elements

1. Introduction

Long terminal repeats (LTRs) retrotransposons are major components of a plant genome. For instance, the grass genome contains around 14% LTR retrotransposons; the maize genome contains 50–80% retrotransposons and barley genome contains at least 70% retrotransposons (McCarthy et al., 2002; McCarthy & McDonald, 2004; Zhang & Wessler, 2004). According to another work, the human and mouse genomes contain 8% LTR retrotransposons and 10% LTR retrotransposons (Li et al., 2001; Zhang & Wessler, 2004), respectively. LTRs are also important structures in retroviruses and endogenous retroviruses. They have LTRs at either end, which play a role in the transposition process (Brown, 1999). Besides, several works have shown that LTR retrotransposons cause some mouse diseases (Kazazian, 1998).

Long terminal repeats and degenerated LTRs are important and abundant components in eukaryotic genomes. LTRs have an important impact on genomic functions and structures. Using the conventional approach to search for LTRs is time-consuming and labor intensive so developing a system that can detect and classify the LTRs is important.

In a recent investigation, two tools were used to detect LTR retrotransposons. They were LTR_STRUC (McCarthy & McDonald, 2003) and RepeatMasker (Smit, 1993). Another investigation mentioned that profile hidden Markov models (HMMs) represent a novel approach to detect transposable elements (TEs) (Juretic et al., 2004).

[*] Corresponding author. Address: Department of Computer Science and Information Engineering, National Central University, Taiwan.
E-mail address: horng@db.csie.ncu.edu.tw (J.-T. Horng).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2007.10.045

LTR_STRUC (McCarthy & McDonald, 2003) is an approach that can find new LTR families. LTR_STRUC successfully identified 12 families of *Caenorhabditis elegans* LTR retrotransposons and 32 new families of LTR retrotransposons in the rice genome (McCarthy & McDonald, 2003). LTR_STRUC uses four structural features to identify LTRs. These structural features are primer binding sites, the polypurine tract, the dinucleotides ends of each LTR and LTR insertion sites. The algorithm was based on the fact that the LTRs are duplicated in the flanking LTR retrotransposons. LTR_STRUC finds an initial pair of matches and extends the pairs by alignment. After a pair of matching regions is detected, the program identifies the LTR endpoints.

RepeatMasker (Smit, 1993) is a tool that can identify the interspersed repeats and the low-complexity sequences. The RepeatMasker is always used with a search engine and a sequence database. The sequence database is always RepBase Update (Jurka, 2000) and the search engine is crossmatch, which is implements the Smith–Waterman–Gotoh algorithm developed by Phil Green (unpublished data). The Repbase Update contains about 1840 transposable elements.

ModelGenerator (Frech et al., 1997) is a program for modeling mammalian and avian C-type LTRs, Lentivirus LTRs and B1 elements. This method has two parts – model generation and model recognitions. Each generated model is composed of various components. ModelInspector identifies LTR elements based on this information. They are consensus elements, matrix elements, hairpins, direct repeats, short multiple repeats, terminal repeats, element classes and regions.

Some functional and significant regions in LTRs are conserved in most instances of the family. This investigation presents a data-mining method for learning significant conserved regions as fingerprints of LTRs. These fingerprints can be used efficiently to find LTRs but also to evaluate the specificity, sensitivity and accuracy of RepeatMasker, LTR_STRUC and the proposed method. These novel LTRs are useful in analyzing the distribution of LTR retrotransposons and the relationship between LTR retrotransposons and evolution.

2. Materials and methods

This study proposes a data-mining system, which identifies and analyzes LTRs in the human genome. Scanning the human genome is scanned by RepeatMasker to collect the training dataset. Finding motifs and clustering motifs generates the profiles of each LTR family. Putative LTRs are detected by scanning the genomic sequences with profiles, and classified into LTR families. Fig. 1 presents an overview of the system flow.

2.1. Data-preprocessing phase

RepeatMasker is a tool used to mask transposable elements based on RepBase Update. RepBase 8.2 includes 624 human TE consensus sequences, in 130 LTR families. The source of the human genome sequence is obtained using Ensembl 19.34a. Ensembl (Hubbard et al., 2002) is a joint project between EMBL-EBI and the Sanger Institute to develop a software system that automatically annotates eukaryotic genomes. Ensembl contains 26,614 contigs and a total of 2.84 billion nucleotides distributed in 24 chromosomes. RepeatMasker detected 12,232 LTR sequences in the human genome sequence.

2.2. Training phase

For each LTR family, about 100 sequences were chosen and five motifs, five conserved regions among 100 sequences, were found using MEME (Bailey & Elkan, 1994). Twenty-five motifs were found by performing the above process five times. However, some of the motifs are highly similar to each other and are redundant for LTR detection. Similar motifs are combined into motif clusters based on CompareACE (Hughes et al., 2000) and *K*-mean clustering (Han & Kamber, 2001). CompareACE is a scoring method based on the Pearson correlation coefficient between two motif alignments. *K*-mean clustering is an algorithm for clustering or grouping motifs into clusters by considering CompareACE scores. The centroids of the motif clusters are chosen to represent the group of motifs and profile HMMs are constructed.

2.3. Detection phase

First, these profile HMMs are used to scan genomic sequences. HMMER (Eddy, 1998) returns both a score and an *E*-value. The HMMER bit scores are the base-two logarithm of the ratio between the probabilities that the query sequence is a significant match to the probability that a null model matches it. The *E*-value indicates the expected number of false positives at a given bit score. The *E*-value measures the statistical significance of the bit score, which indicates how well the sequence matches the HMM. Therefore, HMMER bit scores are used as the scoring function herein.

For each sequence of the input sequences, all possible regions of motif hits are generated. In Eq. (1), $M_{i,j}(F_k)$ are the motifs in a region between positions i and j and belong to the LTR family F_k

$$M_{i,j}(F_k) = \{x | x \in F_k, i \leqslant P(x) \leqslant j\} \quad (1)$$

where x is the motif that belongs to F_k, which is defined in RepBase, and $P(x)$ is the position of motif x. Each region will be treated as a candidate LTR, if the region contains at least one motif. Each candidate LTR is classified into one LTR family, if the candidate LTR satisfies the following limitations.

(1) $M_{i,j}(F_k)$ is the largest for all families between positions i and j. If more than one maximum motif set

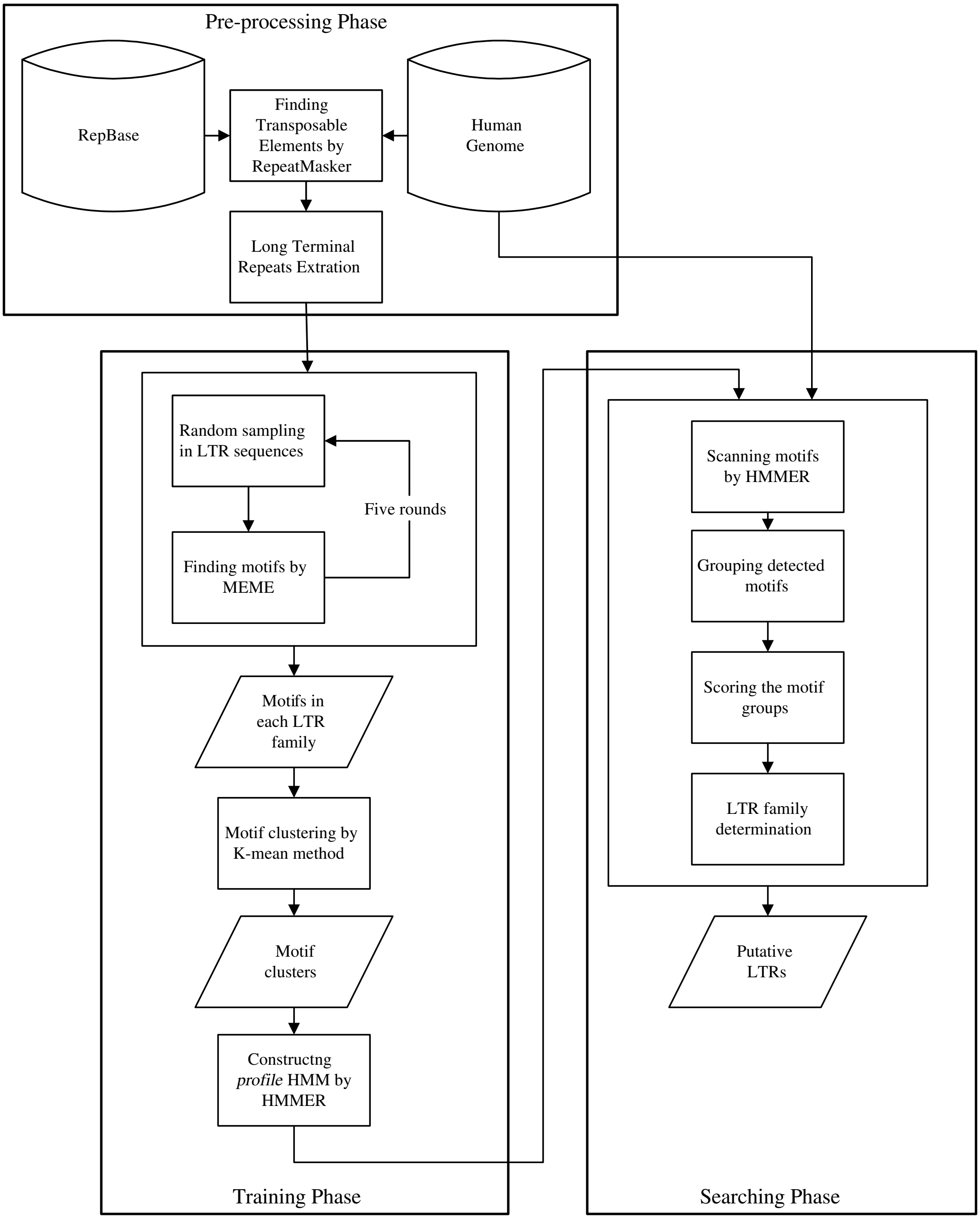

Fig. 1. Overview of system flow.

exists, the one with the largest score is selected. The scoring equation is

$$score(M) = \sum_{m \in M} hmmscore_m \tag{2}$$

where M is one of the maximum motif sets, and $hmmscore_m$ is the HMM bit score of motif m.

(2) The length of candidate LTR (i.e. j–i) must be less than or equal to $length\ (F_k)$.
(3) The $score(M)$ of each motif set must be greater than or equal to $score_{lb}$, where $score_{lb}$ is a user-defined threshold. (see below.)

2.4. Selecting optimal threshold

The detected LTRs are compared to the result of RepeatMasker to maximize the accuracy. The characteristics of LTR families differ, so the most accurate threshold for each LTR family is sought and the threshold of $score_{lb}$ is determined. All data are selected from each family as true data and the sequences are randomly sampled from the other families as false data (Table 1). Each dataset contains equal numbers of true and false data. The experimental dataset herein is tested in 15 rounds and the average sensitivity and specificity are calculated to revise the threshold $score_{lb}$.

The accuracy is determined for each combination and the thresholds at maximum *F_measure* are determined. Fig. 2 presents the algorithm used to find the threshold. Fig. 3 presents the results of the model evaluation and the average *F_measure* is 0.94.

3. Results

A total of 1077 profiles were built after the motif set was clustered and each family contains about 8.2 non-redundant motifs.

An all-against-all comparison of 1077 profiles is performed using the program CompareACE to assess the specificity of the whole profiles. For each profile, the most similar motif in other profiles, except those in the same family, are sought. Fig. 4 reveals that the *X*-axis represents the percentage of all 1077 motifs and the *Y*-axis represents the similarity of the most similar motif. About 75% of the motifs are family-specific and about 25% are similar to those of the other families. This result reveals that most profiles can be used to classify putative LTRs into correct families.

3.1. Datasets for detection

Chromosomes 21 and 22 are placed in the first and second dataset, respectively. 45 Mb, or ~1.8% of the human genome, is also randomly sampled from the Ensembl database, as the third dataset given in Table 2.

3.2. Comparison between LTR_STRUC and RepeatMasker

Several experiments were designed to reveal the differences among the proposed approach, LTR_STRUC and RepeatMasker. Initially, the LTRs were detected by RepeatMasker and treated as real LTRs. Each putative LTRs detected by the proposed approach were classified into three groups.

- *Overlap*: LTRs were successfully found by both methods.
- *Novel*: LTRs were successfully found by LTR_STRUC but not by RepeatMasker.
- *Lost*: LTRs were found only by RepeatMasker.

The LTR_STRUC approach is based on searching for a pair of similar regions. Therefore, LTR_STRUC and RepeatMasker yield quite different results. The following comparison is made to show the differences between them.

```
for each LTR family
    for $score_{lb}$ = 10 to 40
        for each random sample
            execute LTR detection with threshold $score_{lb}$
            calculate sensitivity and specificity
        next sample
        calculate average sensitivity and specificity
        calculate $F\text{-}measure$
    next $score_{lb}$
    report $score_{lb}$ with maximum $F\text{-}measure$
next family
```

Fig. 2. The algorithm is to find the optimal threshold of *score*$_{lb}$.

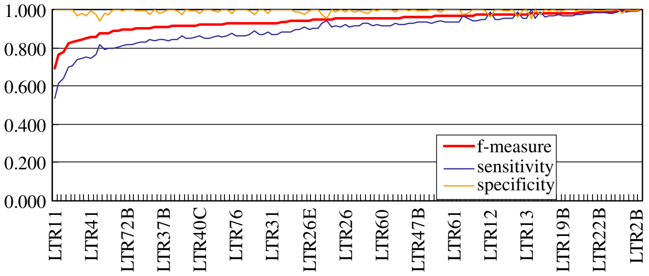

Fig. 3. Sensitivity, specificity and f-measure of the optimal threshold.

All datasets were scanned by LTR_STRUC with default parameters. The overlapping groups are the LTRs that are detected by both methods.

Table 3 compares LTR_STRUC and RepeatMasker. In the first dataset, two (2.5%) LTRs were detected by LTR_STRUC and RepeatMasker. A total of 1690 (99%) LTRs were detected only by RepeatMasker. In the second dataset, eight (7.2%) LTRs were detected by LTR_STRUC and RepeatMasker. A total of 1156 (99%) LTRs were detected only by RepeatMasker. The LTR_STRUC found only a few LTRs, totaling 10% of the LTRs found by RepeatMasker. About 0.6% of the LTRs were identified by both approaches.

3.3. Comparison with RepeatMasker

Initially, RepeatMasker detected LTRs, which were treated as real LTRs. Each putative LTR detected by the proposed approach was classified into one of four groups.

- *Consistent*: The LTRs were found and classified into correct families.

Table 1
Evaluation datasets for family A

Datasets	Source	Description
Positive samples	All family A LTRs	Sequences that were classified into family A by RepeatMasker
Negative samples	All LTRs except family A	Sequences that were classified into other families and randomly sample to the same size of positive samples

Table 2
Datasets for detection

Dataset	Source	# of configs	Total length (Mb)	Description
HUMAN_21	Ensembl Database	485	34	Human chromosome 21
HUMAN_22	Ensembl Database	539	32	Human chromosome 22
HUMAN_RND	Ensembl Database	500	45	Randomly sample 500 human contigs from Ensembl Database

Table 3
Comparison between LTR_STRUC and RepeatMasker

Dataset	# Found by RepeatMasker	# Found by LTR_STRUC	Overlap	Novel	Lost
HUMAN_21	1692	80	2	78	1690
HUMAN_22	1164	110	8	102	1156
HUMAN_RND	1877	136	8	128	1871

- *Inconsistent*: The LTRs were found but were classified into inappropriate families.
- *Novel*: The LTRs were found by the proposed approach but not by RepeatMasker.
- *Lost*: The LTRs were found only by RepeatMasker.

The datasets for each experiment are those constructed in the above section and the statistics are given in the following section.

Table 4 compares the proposed approach and RepeatMasker. In the first dataset, 1457 LTRs were detected by both the proposed approach and RepeatMasker. The family classifications of 272 (19%) LTRs are differ between the proposed approach and RepeatMasker. A total of 218 (11%) LTRs were detected only by RepeatMasker.

In the second dataset, 963 LTRs were detected by the proposed approach and RepeatMasker. The family classifications of 189 (19%) LTRs differed between the proposed approach and RepeatMasker. A total of 157 (13%) LTRs were detected only by RepeatMasker. Table 5 presents the top five families in the inconsistent group. A comparison the results depicted in Fig. 4 demonstrated that the differences between classifications were caused by the common motifs.

As indicated by Table 6, the top five families are in the novel group. The number of LTRs that belong to LTR66 greatly exceeds the numbers in other families because two of the LTR66 motifs were also present in L1MC4, MER63C and MER63D. (See Discussions in Section 4.)

In the experiments herein, about 90% of the LTRs were identified by both RepeatMasker and the proposed approach. Basically, about 63% LTRs of the results of the proposed method are novel. The results of these experiments reveal that most of the LTRs detected by

Table 5
Top five families in the inconsistent group

LTR family	Consistent	Inconsistent	Novel	Lost
LTR13A	0	16	0	0
LTR2	5	16	0	0
LTR28	5	16	39	2
LTR12	17	15	1	0
LTR8	51	15	10	5

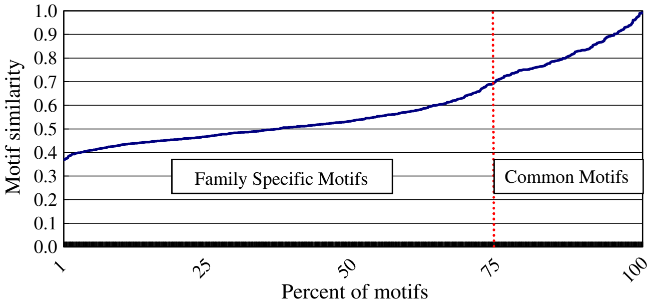

Fig. 4. Searching for the most similar motif in other families.

Table 6
Top five families in the novel group

LTR family	Consistent	Inconsistent	Novel	Lost
LTR66	3	1	1058	3
LTR48B	10	6	120	1
LTR33A	30	13	102	6
LTR37A	26	6	92	5
LTR65	5	0	88	1

Table 4
Comparison with RepeatMasker

Dataset	# Found by RepeatMasker	# Found by our approach	Overlap		Novel	Lost
			Consistent	Inconsistent		
HUMAN_21	1692	4090	1185	272	2633	218
HUMAN_22	1160	2538	774	189	1575	157
HUMAN_RND	1876	4405	1321	305	2779	208

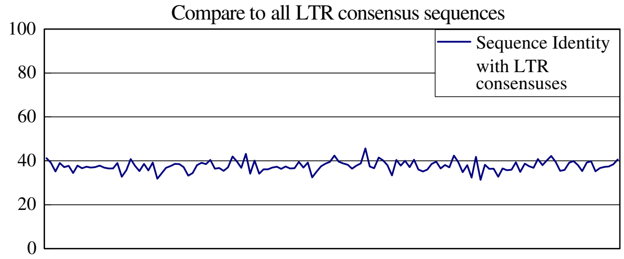

Fig. 5. Sequence identity between LTR consensuses and AJ511661.

RepeatMasker were covered by the proposed approach. Besides, the motif profiles can correctly classify LTRs into families.

3.4. Case study: HERV15 recombination

A case of real recombination was tested using the proposed method to prove its ability to detect the special recombination of LTRs. It is a HERV recombination in human Y chromosome. The NCBI accession number of this case is AJ511661.

Table 7
HMMsearch result of LTR66-0 in RepBase

Sequence	Start	End	HMM bit score	*E*-value
LTR66	12	51	32.6	9.30E−08
LTR66	397	436	26.1	8.80E−06
L1ME4	367	406	15.8	0.011
LTR66	500	539	15.2	0.017
L1MC4	1375	1414	12.4	0.041
L1MC3	1351	1390	10.1	0.073
MER63C	40	79	10.0	0.075
MER63D	40	79	10.0	0.075

Table 8
HMMsearch result of LTR66-7 in RepBase

Sequence	Start	End	HMM bit score	*E*-value
LTR66	114	153	32.1	1.40E−07
LTR66	397	436	30.9	3.20E−07
L1MC4	1377	1416	20.7	0.00036
MER63C	38	77	16.2	0.0084
MER63D	38	77	16.2	0.0084
MER4BI	1160	1199	13.1	0.069

RepeatMasker cannot detect this particular case and Fig. 5 reveals the local alignment score with all LTR consensuses. The proposed approach can detect and classify it into the family LTR48B. The order of motifs in this case and the complete structure differs greatly. The detection method of RepeatMasker is based on sequence identity; therefore, RepeatMasker cannot detect this case, which can be detected using LTR fingerprints and the proposed approach.

4. Discussions and conclusions

This work develops a novel method of LTR detection. LTR sequences can be analyzed to find some specific motifs as LTR fingerprints and built them into HMM profiles. The experimental results indicate that the proposed approach can not only discover most of the LTRs found by RepeatMasker, but also detect some novel LTRs. Additionally, the novel LTRs may be structurally recombinant or degenerated.

The main difference between the proposed approach and the traditional approach, such as BLAST or RepeatMasker, is that the former searches for the most conserved region as the profile. Besides, the profiles can used to classify LTRs into the correct family. Restated, the LTRs with particular combination of fingerprints were be detected successfully.

About 63% of the novel LTRs were classified into LTR66 and most were detected by motif LTR66-0 and LTR66-7. Hence, LTR66-0 and LTR66-7 are searched in all consensus sequences defined by RepBase. Tables 7 and 8 show that L1MC4, MER63C and MER63D contain a similar region to those in LTR66-0 and LTR66-7. Accordingly, these motifs may be significant regulatory sites in these families.

Table 9 compares the LTR detection tools. RepeatMasker links fragments that belong to a single transposable element, but it does not consider the structure. LTR_STRUC reports the putative primer binding site and the polypurine tract and open reading frames in each LTR. ModelGenerator generates a consensus model from regulatory units, direct repeats and hairpins in specified LTRs.

Table 9
Comparison among the LTR detection tools

	RepeatMasker	LTR_STRUC	ModelGenerator/ ModelInspector	Our approach
Materials	RepBase	None	Collected sequences	RepBase
Characteristics	Search for similar regions with conensus sequences	Search for a pair of similar regions as LTR	Generate consensus model using collected sequences	Finding significant conserved motifs and detect LTRs based on the motifs
Searching algorithm	Smith-Waterman-Gotoh algorithm	Sequence comparison	ModelInspector	HMMER
Degenerated LTRs detection	Normal	Few	Few	A lot

Juretic et al. (Jurka, 1998) annotated transposable elements in the rice genome by establishing and scanning HMM profiles of the Mutator-like element and the miniature inverted-repeat transposable element superfamily. They asserted that profile HMMs could support RepeatMasker and improve its capacity to detect degenerated copies of TEs.

The scoring mechanism of the proposed approach considers only the similarity between fingerprints and the input sequence. Therefore, the ordering or topology of motifs in each family can be applied in the scoring mechanism. A higher score implies greater similarity between structures. Common/specific status can be considered in the scoring function.

LTR is the major characteristic of the LTR retrotransposon and the profiles of the other internal features such as gag, pol and env gene can be established. Each feature of the LTRs with complete structures can be identified precisely, and novel LTRs are useful in analyzing the distribution of LTR retrotransposons. This study focuses on human LTRs and can be effectively applied to other species.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC96-2311-B-008-004-MY2 and NSC96-2221-E-008-104-MY3.

References

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference Intelligent System for Molecular Biology, 2*, 28–36.

Brown, T. A. (1999). In F. Kingston (Ed.), *The repetitive DNA content of genomes* (pp. 138). New York: John Weley.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755–763.

Frech, K., Danescu-Mayer, J., et al. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *Journal of Molecular Biology, 270*(5), 674–687.

Han, J., & Kamber, M. (2001). *Clustering analysis. Data mining concepts and techniques*. Morgan Kaufman, 335.

Hubbard, T., Barker, D., et al. (2002). The Ensembl genome database project. *Nucleic Acids Research, 30*(1), 38–41.

Hughes, J. D., Estep, P. W., et al. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology, 296*(5), 1205–1214.

Juretic, N., Bureau, T. E., et al. (2004). Transposable element annotation of the rice genome. *Bioinformatics, 20*(2), 155–160.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Current Opinion Structural Biology, 8*(3), 333–337.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genetics, 16*(9), 418–420.

Kazazian, H. H. Jr., (1998). Mobile elements and disease. *Current Opinion Genetic and Development, 8*(3), 343–350.

Li, W. H., Gu, Z., et al. (2001). Evolutionary analyses of the human genome. *Nature, 409*(6822), 847–849.

McCarthy, E. M., Liu, J., et al. (2002). Long terminal repeat retrotransposons of Oryza sativa. *Genome Biology, 3*(10), RESEARCH0053.

McCarthy, E. M., & McDonald, J. F. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics, 19*(3), 362–367.

McCarthy, E. M., & McDonald, J. F. (2004). Long terminal repeat retrotransposons of Mus musculus. *Genome Biology, 5*(3), R14.

Smit, A. F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research, 21*(8), 1863–1872.

Zhang, X., & Wessler, S. R. (2004). Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. *Proceedings of the National Academy of Sciences of the USA, 101*(15), 5589–5594.