Università di Pisa
Dipartimento di Informatica

Technical Report: TR-09-15

About LTR recurrence in yeast strains

Giulia Menconi [a], Claudio Felicioli [b] and Roberto Marangoni [c,d]

[a]Istituto Nazionale di Alta Matematica, Roma, Italia
[b] Istituto di Scienza e Tecnologie dell'Informazione, CNR, Pisa, Italia
[c]Dipartimento di Informatica, Università di Pisa, Italia
[d]Istituto di Biofisica, CNR, Pisa, Italia

Address: Largo B. Pontecorvo 3, 56127 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

About LTR recurrence in yeast strains

Giulia Menconi [a], Claudio Felicioli [b] and Roberto Marangoni [c,d]

[a]Istituto Nazionale di Alta Matematica, Roma, Italia
[b] Istituto di Scienza e Tecnologie dell'Informazione, CNR, Pisa, Italia
[c]Dipartimento di Informatica, Università di Pisa, Italia
[d]Istituto di Biofisica, CNR, Pisa, Italia

Abstract

Here we report a preliminary analysis of transposon long terminal repeats (LTR) recurrence within several yeast (*Saccharomyces cerevisiae*) strains.

1 Introduction

Modern representations of the genomes of higher organisms enclose some mobile elements (ME), that are not associated with a defined position along the chromosome, but can move with peculiar dynamic properties. The most important class of mobile entities is represented by the transposons (also transposable elements, TE): mobile elements in the DNA, which are generally considered an example of "selfish DNA" sequence, only partially linked to the host genome. Although extremely different from each other for the specific biological information they can carry on, they share the ability to multiply and invade the genome of various species, with variable results ranging from lethal insertions to neutral or even adaptive effects. Discovered in the maize by Barbara McClintock [9] in the 1950, as of today it is known that they represent a very consistent part of the higher organisms genomes (about 15% in the human genome). Transposons are elements of variable length, that can move around the genome by randomly jumping from a locus to another, or even to a different chromosome. They may contain genes, promoters and non-coding sequences. When they relocate to a new locus in the genome, this movement can generate consequences of variable and unpredictable importance: they can destroy the transcription of a gene, as well as they can activate an unfunctional gene.

The large quantity of these transposons, and the fact that they can interact with each other, by inhibiting or enhanching their duplication, and the evidence that they can interact with standard genes of the host organism (they act in a parasite-like way) have lead to a contemporary representation of genomes as "ecosystems", while "resident" genome and mobile elements compete for chemical resources [5]. The genome stability is the result of a continuous interaction between the TE and the *not*-TE components.

This work is focused on *Saccharomyces cerevisiae* genome, as yeast molecular and cell biology has accumulated a great amount of qualitative and quantitative data of diverse cellular processes (see for instance [13]). We shall

focus on Ty1 transposon [11] and extract information from the data concerning the insertion and action of TE [3]. This should also comprehend the behaviour in different strains [10, 2].

More in details, we shall analyse the linguistic recurrence of the long terminal repeats (LTR) in Ty1, which are flanking segments of transposon sequences. We shall perform a comparative analysis among a selected group of a wider collection of complete genomes of different strains of *S. cerevisiae*, newly sequenced and introduced in [8].

2 Materials and methods

2.1 Yeast

Yeast (*Saccharomyces cerevisiae*) is a unicellular organism, widely used as an experimental system for molecular biology since around 1960. The reference (RefSeq) genome is around 12.8 Mb and it is organised in 16 chromosomes (size ranging from 250 kb to more than 2500 kb) and a mithocondrion. After sequencing, around 6000 ORFs have been identified, most of wich are likely to encode specific proteins.

In 2009, Liti et al. reported nearly complete genome sequences of *S. cerevisiae* and *S. paradoxus* from a large variety of sources and locations [8]. Their work was aimed at studying variations in gene content, single nucleotide polymorphisms, nucleotide insertions and deletions and so on. In particular, we worked on the ABI data for 39 *S. cerevisiae* strains, which were submitted to the NCBI Trace Archive and are available through the yeast genome resequencing project website [12]. We denote that collection by "Liti's strains".

The new assemblies are lacking some precise base identification, that varies among chromosomes (the so-called N-content, Fig. 1, left). The new assemblies are also variable in chromosome length (average, min length, max length), see Fig. 1 (right).

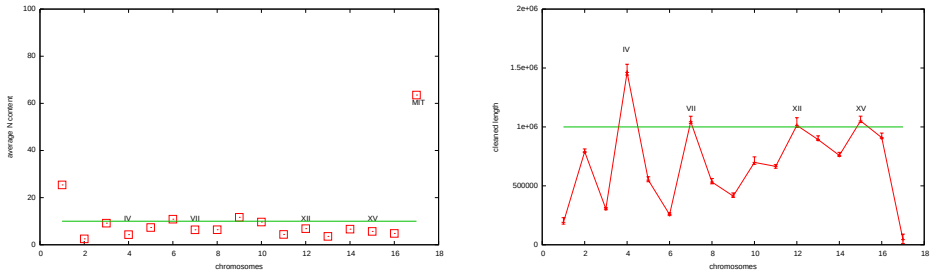


Figure 1: *Content of unidentified nucleotide on each chromosome (left) and length (min, max, mean value) of indentified sequences (right). For both measures the averages are taken on all Liti strains.*

2.2 LTR retrotransposons

Yeast retrotransposons are Long Terminal Repeats (LTR) retrotransposons [7]. Typically, they are 6.3 kb long and the last 330 bp at each end are direct

repeats (δ units). They typically have two genes (2 ORFS: TyA and TyB). Such ORFs are expressed in the same direction, but read in different phases and overlapping by 13 aminoacids; the organisation and functions of TyA and TyB are analogous to the behaviour of retroviral *gag* and *pol* functions. Ty elements are classic retrotransposons, using an RNA intermediate: an intron is inserted into an element to generate a unique Ty sequence, that is placed under control of a GAL promoter on a plasmid and introduced into yeast cells. Transposition results in the appearance of multiple copies of the transposon in the yeast genome, all lacking the intron.

During reverse transcription, a retrotransposon generates a copy of itself that is integrated into a new location in the genome. Once it has been inserted, a new element may experience 3 fates: transposition, excision, mutation. Transposition: generates another identical copy (as efficient as the mother one). Mutation: either in LTRs or in the two flanked regions. Over evolutionary time, as the $5'$ and $3'$ LTRs within a single element gradually diverge, the degree of sequence divergence between them can provide an estimate of the time elapsed since the element inserted. Excision: the element and one LTR is exceised and permanently lost, while the other LTR remains as *solo LTR* marker of the site from which the element was lost.

Therefore, the LTR recurrence within yeast genomic DNA is of interest in order to investigate Ty dynamics.

There are around 30 copies of Ty1 type in a RefSeq yeast genome; in addition, there are around 100 independent δ units, called solo δs.

We chose to focus on Ty1 LTR in Chromosome IV (which is around 1.5×10^6 bp long in RefSeq) which is the longest chromosome and whose N-content is around 4.3%, on average on Liti's strains.

Available Ty1 genomic sequences for RefSeq have been downloaded from SGD (Saccharomyces Genome Database [13]), they belong to two subclasses: YDRC and YDRW.

- YDRCTy1-1: Chr IV 651417-645500, rev. compl., 5918 nt
- YDRCTy1-2: Chr IV 884218-878301, rev. compl., 5918 nt
- YDRCTy1-3: Chr IV 992639-987147, rev. compl., 5493 nt
- YDRWTy1-4: Chr IV 1095765-1101690, 5926 nt
- YDRWTy1-5: Chr IV 1206697-1212614, 5918 nt

Such Ty1 sequences are extremely similar to eachother (BLAST alignment score $97 - 98\%$). Dissimilarities are concentrated on LTRs.

On RefSeq chromosome IV, 27 different Ty1 LTR genomic sequences are available.

Among them, we tested our methodology on some randomly selected LTR sequences:

- YDLCdelta1 (317 nt)
- YDRCdelta2 (273 nt)
- YDRWdelta10 (242 nt)
- YDRWdelta20 (332 nt)

We perfomed our comparative analysis on five strains, with a different geographic origin: NCYC110 (Africa), Y55 (Europe) and S288c, SK1 and YPS128 (Americas).

2.3 n-grams on LTR

For each LTR genomic sequence, we created the n-gram collection of lengths n from 10 to around the length of the LTR: they are the n-long overlapping segments within the LTR sequence.

If L is the LTR total length, then there are $L-n+1$ different windows, each one containing one of the 4^n possible n-grams over the $\{A, C, G, T\}$ alphabet.

For each n-gram found, its occurrences over the complete chromosome IV sequence of each strain have been computed and localized, after a suffix tree has been generated on the chromosome.

We remark that, focusing on $n \leq 30nt$, the n-grams in any two windows are pairwise different, with the only exception of only one 10-gram in YDRWdelta20. This means that the n-gram structure is extremely variable and therefore the localization of recurrent n-grams along the chromosome could suggest how and where the transposition action has been either successful (almost intact LTR) or has failed (only fragments remained).

3 Results

We show some preliminary observations about the n-gram analysis on LTR YDLCdelta1 recurrence over chromosome IV in strains Y55, NCYC110 and YPS128.

As n grows, recurrent LTR fragments are clustered to a few regions of LTR complete sequence. This is shown for instance for Y55 strain in Fig.2. For each n-gram length (x axis), every n-gram occurrence is screened on the z axis, following the order of appearance of the specific n-gram within the LTR sequence (y axis).

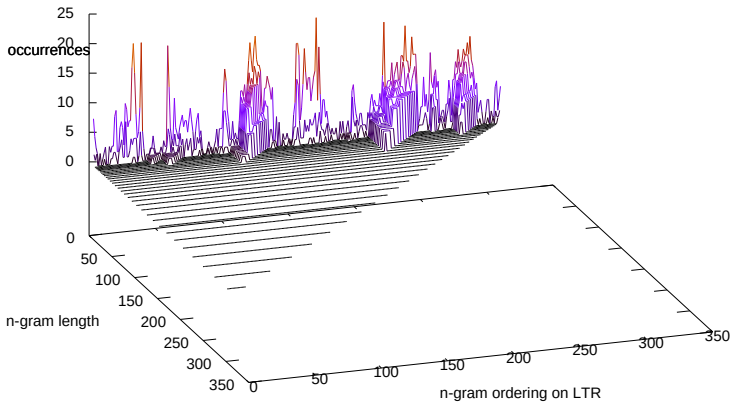
For small values of n, n-grams are easily found within all the chromosome sequence, but for $n \geq 10$ only a few subsequent overlapping n-grams occur in a notable way. We recall that as LTR sequences are around $3 \times 10^3 bp$ long, then there may be a statistically meaningful number of different words of length n only for $n \leq 4$ since $4 \leq \log_4(3 \times 10^3) \leq 5$. Therefore, any result concerning words longer than $n = 5$ concerns meaningful words: their recurrence is not due to mere statistical motivations.

The recurrence of those clustered n-grams may vary from strain to strain. Now, the challenge is about finding a fruitful way to visualize those data.

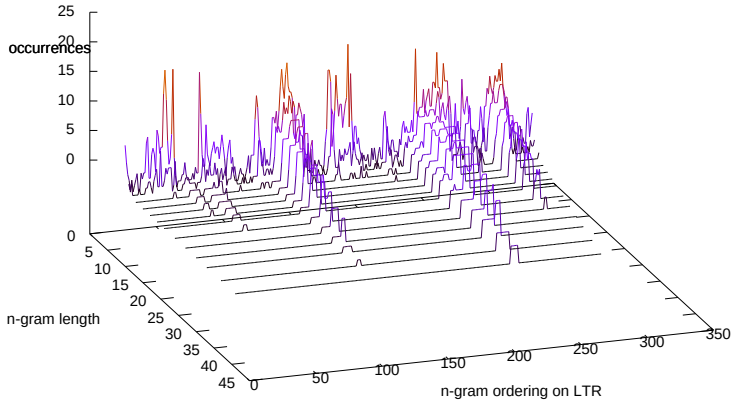
When analysing in details where each specific n-gram is located within the chromosome IV of each strain, we noticed that the same pattern may be extremely repeated in one strain and be more rare in other strain.

In Figure 3, we show what the fate is for 17-gram $ATTGATAATGTAATAGG$ from LTR YDLCdelta1 in strains Y55, NCYC110 and YPS128.

Plot (a) compares Y55 and NCYC110: the 8 occurrences of the 17-gram in NCYC110 are co-localized in Y55, even if in Y55 there are 3 more occurrences that are anyway clustered to previous ones.

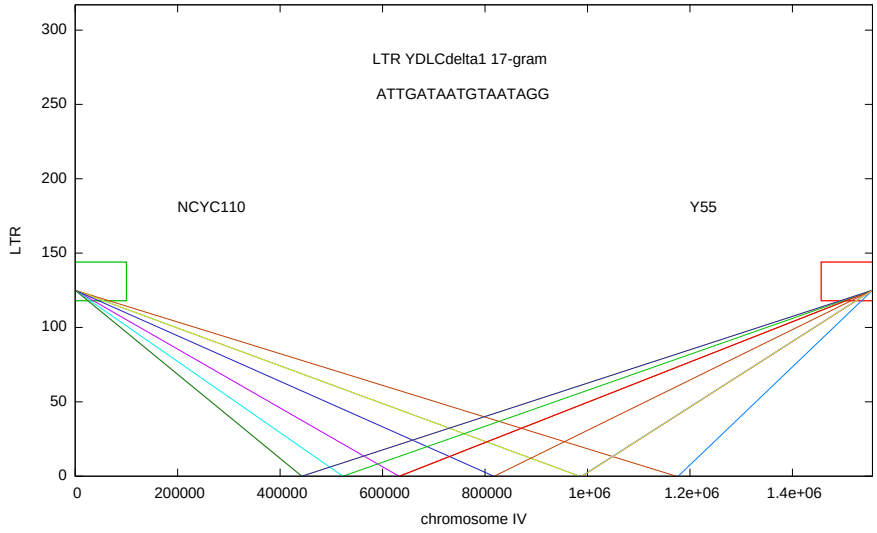


(a)

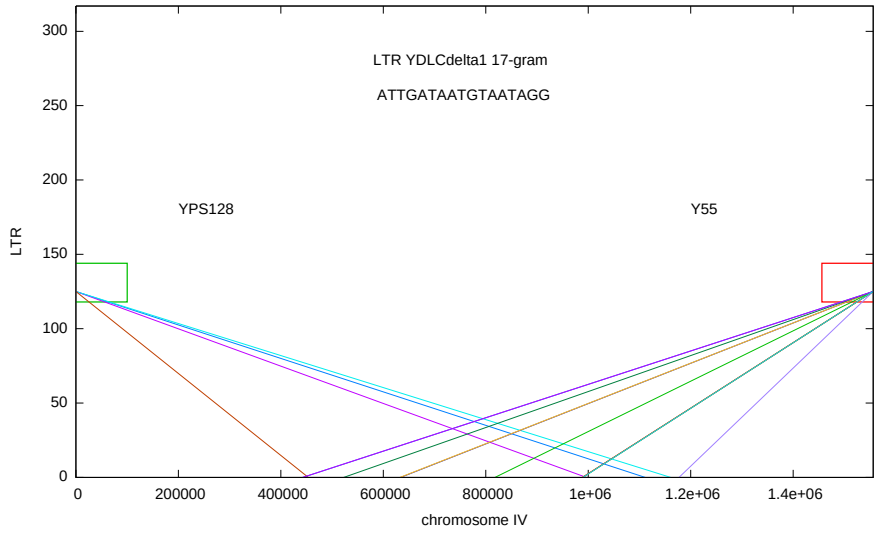


(b)

Figure 2: *LTR YDLCdelta1 n-gram occurrences within Y55 strain: n-gram length from 1 to 300 (a) and a zoom for length up to 45 (b).*



(a)



(b)

Figure 3: *Localization of 17-gram* $\boxed{ATTGATAATGTAATAGG}$ *within chromosome IV in Y55, NCYC110 (a) and YPS128 (b) strains.*

The same is not true for Y55 and YPS128: plot (b) shows that the 4 occurrences in YPS128 are almost co-localized with relatively close regions in Y55 chromosome IV, but other regions in YPS128 do not present analogies with Y55.

These preliminary remarks on LTR recurrences suggest many questions. May we extend recurrence from n-grams to variable length segments? Why some regions within the chromosome seem to be a preferred target for LTR fragment recurrence? Is the localization a conserved property within strains of the same origin? May we build an LTR-based phylogeny and does it compare to whole-genome phylogeny?

In the future, we shall improve the visualization tool in order to understand the evolution of LTR fragment recurrence over the chromosome. As a further step, we plan to build a tool based on the so-called masks as in [1] (motif discovery with "don't care" symbols) in order to extract information about the collection of repeated structures in common among Liti's strains.

Acknowledgment

The work of G.M. was supported by a post-doc research scholarship "Compagnia di San Paolo" awarded by the Istituto Nazionale di Alta Matematica "F. Severi".

References

[1] Battaglia G., Cangelosi D., Grossi R., Pisanti N., Masking patterns in sequences: a new class of motif discovery with don't cares, Theor. Comput. Sci., doi: 10.1016/j.tcs.2009.07.014 (2009)

[2] Braiterman L.T., Monokian G.L.,Eichinger D.J.,Merbs S.L., Gabriel A., Boeke J.D. (1994) In-frame linker insertion mutagenesis of yeast transposon Tyl : phenotypic analysis, Gene, vol. 139, no1, pp. 19-26

[3] Horecka J.,Jigami Y.(2000) Identifying tagged transposon insertion sites in yeast by direct genomic sequencing, Yeast, 16,10:967-970

[4] Klipp E.(2007) Modelling dynamic processes in yeast, Yeast, 24:943-959

[5] Le Rouzic A., Dupas S., Capy P. (2007) Genome ecosystem and transposable elements species, Gene 390, 214-220.

[6] Le Rouzic A., Boutin T.S., Capy P. (2007) Long-term evolution of transposable elements, PNAS 104, 49:19375-19380.

[7] Lewin B, Genes IX, Jones and Bartlett Publishers, (2008)

[8] Liti G. et al, Population genomics of domestic and wild yeasts, *Nature*, **458** (2009)

[9] Mc Clintock B.(1950) The origin and behavior of mutable loci in maize, PNAS, 36: 344-355.

[10] Suzuki C.,Hori Y.,Kashiwagi Y.(2003)Screening and characterization of transposon-insertion mutants in a pseudohyphal strain of Saccharomyces cerevisiae, Yeast, 20, 5: 407-415

[11] Wu X., Jiang Y.W (2008)Overproduction of non-translatable mRNA silences. The transcription of Ty1 retrotransposons in S. cerevisiae via functional inactivation of the nuclear cap-binding complex and subsequent hyperstimulation of the TORC1 pathway, Yeast,25,5:317-347

[12] www.sanger.ac.uk/Projects/S_cerevisiae

[13] www.yeastgenome.org