Prediction of transposable elements evolution using tabu search

Lingling Jin
Department of Computing Science,
Thompson Rivers University,
Kamloops, Canada
Email: ljin@tru.ca

Ian McQuillan
Department of Computer Science,
University of Saskatchewan,
Saskatoon, Canada
Email: mcquillan@cs.usask.ca

***Abstract*—Transposable elements (TEs) are DNA sequences that can move or copy to new positions within a genome. Due to their abundance in many species, predicting the evolution of these TEs within a genome is a major component of understanding the evolution of the genome generally. The sequential interruption model is defined between TEs that occur in a single genome, which has been shown to be useful in previous literature in predicting TE ages and periods of activity throughout evolution. This model is closely related to a classic matrix optimization problem: the linear ordering problem (LOP). By applying a well-studied method of solving the LOP, tabu search, to the sequential interruption model, a relative age order of all TEs in the human genome is predicted in only 38 seconds. A comparison of the TE ordering between tabu search and the previously existing method shows that tabu search solves the TE problem exceedingly more efficiently, while it still achieves a more accurate result. The speed improvements allow a complete prediction of human TEs to be made, whereas previously, ordering of only a small portion of human TEs could be predicted. A simulation of TE transpositions throughout evolution is then developed and used as a form of *in silico* verification to the sequential interruption model. By feeding the simulated TE remnants and activity data into the model, a relative age order is predicted using the sequential interruption model, and a quantified correlation between this predicted order and the input (true) age order in the simulation can be calculated. An average correlation over ten simulations is calculated as 0.738 with the correct simulated answer.**

***Index Terms*—transposable elements, the human genome, evolution, interruptional analysis, tabu search, linear ordering problem**

I. Introduction

Transposable elements (TEs) are one type of repetitive DNA sequences that are found in both eukaryotic and prokaryotic organisms, which have the ability to move or copy to new positions within a genome. TEs are traditionally classified into two broad classes on the basis of their transposition mechanism and sequence organization [1]: Class I elements ("copy-and-paste" mechanism) are those that transpose via reverse transcription of an RNA intermediate, referred to as *retrotransposons*; Class II elements ("cut-and-paste" mechanism) move primarily through a DNA-mediated mechanism of excision and insertion, and are often called *DNA transposons*. The impact of TEs on genome evolution appears to be extensive and they are even believed to promote speciation [2] and can therefore be seen as a driver of evolution. They also can represent a massive fraction of many genomes, from humans (45%) to wheat (>80%). Therefore, the evolutionary history of TE families in a species represents a key component of information regarding the evolution of the genome. Moreover, evidence is emerging that active TEs play a significant role in human biology and disease. They create genetic diversity and integrate preferentially into genes associated with certain functions [3], potentially causing disease. For example, it was found that a molecular mechanism of the Alzheimer's process could be caused by *Alu* elements (one family of TE) leading to dementia [4]. In [5], a method was introduced that could predict the evolutionary history of TEs from a single genome. Using a single genome, rather than requiring sequences from multiple genomes as is commonly required for phylogeny, is useful in situations where many closely related genomes are not available. However, the computational method in [5] was too slow to be able to provide a solution for all human TEs. A method that is efficient enough to not only accurately predict the evolution of all human TEs, but also for any genome, would be important. This is especially significant for large plant genomes that often evolve rapidly, that can have a large TE fraction, and that are less well-studied. As an example, the recent genome assembly of wheat [6] was significantly complicated by the large number of active TEs, and the vast differences from other related genomes. For this reason, it should be efficient enough to be usable on all TEs of very large genomes. This work provides such a method.

A method called interruptional analysis in [5] estimates relative TE ages based on the frequencies with which every TE has inserted itself into every other TE in a genome. It is strategized in two major steps (summarized from [5]): 1. generate an interruption matrix based on the identified pairwise insertion frequencies; 2. generate a TE chronological order using a repositioning method to minimize the sum of non-zero entires above the diagonal, which has the effect

Research supported, in part, by a grant from the Natural Sciences and Engineering Research Council of Canada

Paper published in Proceedings of BIBM 2018, https://doi.org/10.1109/BIBM.2018.8621478.

of ordering the TEs in a hypothetical chronological order of decreasing age (from oldest to youngest). This prediction works because the top-right portion of the matrix represents older TEs interrupting younger TEs, which should be mainly populated by zeros, meaning that there are no interruptions; the bottom-left portion of the matrix represents younger TEs interrupting older TEs, where most non-zero values should appear. Essentially, this interruptional analysis method uses exhaustive search with a worst-case computational complexity of $O(n!)$ total possible orders, which is computational intractable. Though the authors in [5] tried different strategies to decrease the complexity, it is still not practically feasible when the number of TEs in consideration is large. Therefore, they were only able to do so on 405 of the over 1000 human TEs. The interruptional analysis method has the benefit of not relying on any other related genomes. The predictions in [5] were reasonably consistent with traditional phylogenetic analysis, but the authors did not give a complete quantitative assessment.

II. Methods: Linear ordering problem and tabu search for sequential interruption analysis

Previously, we have defined the sequential interruption linear ordering problem [7] — a formal model defined between the interruptions of TEs that occur in a single genome and connected it with the linear ordering problem. In order to describe and compute a solution of the sequential interruption linear ordering problem, a set of matrix rearrangement operations using linear algebra and the original linear ordering problem were also examined. The sequential interruption analysis described in [7] in terms of the linear ordering problem was defined as: given a set of genomic sequences, S, a set of TEs with a fixed ordering on its elements, $\chi_s = \{X_1, X_2, \ldots, X_m\}$, and an interruption matrix of χ_s on S, $IM = [|\Xi_S(X_i, X_j)|]_{\substack{i=1,\ldots,m \\ j=1,\ldots,m}}$ (calculated from the pairwise interruptions that occur in S), the problem is to find a permutation π of χ_s, corresponding to the column and row indices $\{1, \ldots, m\}$, such that the value $f(\pi) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} IM^{(\pi(i),\pi(j))}$ is maximized. Note that in the LOP, the permutation π provides the ordering of both the columns and the rows. The resultant permutation π of χ_s corresponds to a hypothetical chronological order of TE families in χ_s of increasing age.

The LOP is a relatively well-studied problem, known to be $\mathcal{NP}$-hard in computer science [8]; this implies that there likely does not exist a polynomial time algorithm for always calculating an optimal solution (unless $\mathcal{P} = \mathcal{NP}$). Given an interruption matrix of n TEs by the sequential interruption model, an exhaustive search can be used to find the best permutation by applying all possible permutations to the interruption matrix and calculating the sum of the values above the diagonal for each permuted matrix. The ordering of the matrix that achieves the maximum score is the optimal permutation. The exhaustive search algorithm has a complexity of $O(n^2 \times n!)$ (the n^2 does the additions of the upper triangle), which is only usable for very small data sets. Speed is indeed an important issue for predicting TEs ages in this problem because the previous method in [5] was not able to solve it with all human TEs. Heuristic and meta-heuristic methods attempt to find a good, but not necessarily optimal solution to the problem, which is in contrast to exact methods that guarantee to give an optimum solution. Nevertheless, the time taken to search an optimal solution to a difficult problem by an exact method is often much greater than heuristic and meta-heuristic methods. Thus, heuristic and meta-heuristic methods are often used to solve real optimization problems. In the next subsection, one of the meta-heuristic methods, tabu search [9], will be described, then the result of tabu search applied to the sequential interruption model will be provided and compared with the published result in [5].

A. Tabu search and results

Tabu search keeps a table of solutions that are forbidden to guide the search, so that the selection of solutions is limited according to the table of tabu status. Tabu search begins in the same way as an ordinary local search, moving from one solution to another repeatedly until a number of global iterations are performed without improving the best solution found so far. If the search space is seen as a huge set of solutions and only a tiny part of the set can be explored, then tabu search guides the local search process to examine the solution space beyond local optimality. It consists of two search strategies — *intensification* and *diversification* — with complementary objectives to search in the solution set. Intensification favours the exploration of promising areas of the solution space, while diversification moves the search to new regions of the solution space.

Given the interruption matrix, $IM(1015 \times 1015)$, calculated on the human genome *hg38*, the sum of the overall matrix excluding the sum of the diagonal is 381,201. By inputting IM to the tabu search program, a TE ordering that achieves the best superdiagonal score, $f(\pi) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} IM^{(\pi(i),\pi(j))} = 377,417$, is calculated. Since the tabu search is a meta-heuristic algorithm, this score is not guaranteed to be optimal. However, it only took 38 seconds to calculate the best score of this matrix of size 1,015 on MacBook Pro (2.9 GHz Intel Core i5 processor with 16 GB memory). The ability to solve this problem on such a big matrix has made tabu search outperform the method proposed in [5] in terms of efficiency, which was only able to solve a much smaller matrix also without a guarantee of finding the optimal solution.

The resultant ordering from tabu search is then compared with the ordering published in [5] (Giordano et. al.) in two different ways. First, as the method in [5] is compute-expensive, though there were about $1,000$ TEs with interruptions, only 405 were selected for calculating their ordering in the paper. Among these selected 405 TEs, there were 359 of them that are in common with the TEs in the IM that was calculated on the human genome $hg38$. This might be caused by the ongoing updates in Repbase Update [10] during these years, which was required to identify the TEs. The set of the $n = 359$ common TEs are denoted by χ_n. The resultant ordering from tabu

search of χ_n is denoted as π_{tabu}, and the ordering published in Giordano et. al. [5] of these TEs is denoted as π_G. The sub-matrix of IM on χ_n is denoted as IM_{χ_n}. The superdiagonal scores of the two permutations π_{tabu} and π_G on IM_{χ_n} are $f(\pi_{tabu}) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} IM_{\chi_n}^{(\pi_{tabu}(i),\pi_{tabu}(j))} = 165,980$ and $f(\pi_G) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} IM_{\chi_n}^{(\pi_G(i),\pi_G(j))} = 165,591$. Hence, the ordering calculated by tabu search achieves a higher score for this (reduced) data set, which indicates that tabu search is performing better than the method in [5] in terms of the final score.

Second, the similarity between the ordering calculated from tabu search (π_{tabu}) and the ordering published in [5] (π_G) are compared with each other using the Pearson's coefficient of correlation calculated as $\rho = \frac{cov(\pi_G,\pi_{tabu})}{\sigma_{\pi_{tabu}}\sigma_{\pi_G}} = 0.943522$, which shows a strong positive correlation (agreement) between the two orderings (a correlation of 1 means that the two orders are identical, 0 means that there is no correlation, and -1 means negative correlation). The two orderings are then plotted against each other in Fig. 1.

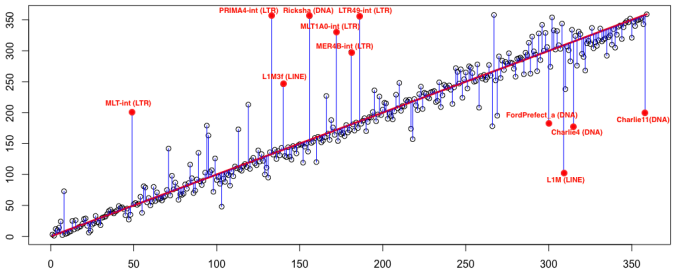

Fig. 1. A comparison between the ordering calculated in [5] (π_G) and the ordering calculated by tabu search (π_{tabu}). The red thick diagonal line represents the case when the two orderings are exactly the same.

As TEs can be active for a period of time, more than one TE can be active at the same time, in parallel. Hence, it is reasonable that the relative age order of these parallelly active TEs are shuffled within a region of the overall ordering, which results in these TEs being "around" the red thick diagonal line in the plot. It can be seen from the figure that most of the elements agree in the two orderings very well as they are very "close" to the red line. There are also some "outliers", which indicate that their positions are distant (very different) in the two orderings. The elements that have a distance of more than 100 positions in the two orderings are marked with their names and types in brackets in Fig. 1.

A preliminary biological analysis and verification of their actual positions (actual evolutionary age) of the elements relative to other TEs was done in [5]. However, another standard technique for verification is to use the help of a simulation, which is created in the next section.

III. TE EVOLUTION SIMULATION

In this section, a simulation is created to imitate a simplification of the evolutionary history of TE propagation. By simulating the TE activities through evolution, the remnants of TEs with their positions in a simulated genome can be generated. Such a simulation is an important tool for understanding transpositions and the evolution of genomes generally. Furthermore, it can be used as a verification tool for TE prediction problems as the ground truth is known. If the known ages and lifespans of TEs from a simulation, and the predicted age order calculated by the theoretical model using the simulated remnants are similar, this serves as an *in silico* verification of the prediction.

The simulation starts from a point in evolutionary time (e.g., 200 million yeas ago or MYA), and simulates the mutations in a genome and the insertions and degrading activities of TEs. As time progresses, TEs are activated when the "current" time matches their input ages. The activated TEs start their transpositional activities while accumulating mutations at the same time. The mutations in TEs decrease the activity levels of these TEs, until they become inactive. The simulation can imitate the activity of the entire lifespan of these TEs in a genome.

A. The `PhyloSim` simulation package

The `PhyloSim` [11] (License: GNU General Public License Version 3) is an object-oriented framework of Monte Carlo simulation (a numerical experimentation technique to obtain the statistics of the output variables of a computational model, given the statistics of the input variables [12]) of sequence evolution written in `R`, which simulates the evolution of DNA or protein sequences by using substitution models of the type of the sequence.

Our simulation of TE transpositions through sequence evolution is written in the `R` language and built on top of the `PhyloSim` package. This is because `PhyloSim` provides functions that simulate random substitutions through sequence evolution, and the TE transposition simulation imitates the replication and insertion of active TEs on a dynamically changing genomic sequence through evolution. Therefore, the `PhyloSim` package provides the necessary generality on which to extend its functionality. Unlike `PhyloSim` that simulates sequence evolution of multiple species under the guidance of a phylogenetic tree, the TE transposition simulation currently only simulates one genomic sequence, as the TE predictions work on a single genome. The workflow of the TE transposition simulation starts from an evolutionary time T and an original genomic sequence. As time is being consumed, mutations are introduced into the genome randomly using the functions provided by `PhyloSim` following a substitution model and the mutation rate; at the same time, TEs are being activated when the current time reaches the input ages of these TEs, and transpositions occur through evolution. This is repeated until time is exhausted. The modified genomic sequence at the end of the simulation represents the current-day genome, which encodes the history of TE activities by their positions of insertions and their interruption patterns, similar to the current-day human genome.

The simulation is subject to a number of parameters such as mutation rate, transposition rates, substitution model etc., and

a set of input data, which are listed and explained in details in the next subsections.

B. Parameters

Some parameters are useful for the simulation as follows.

- Mutation rate (μ): 0.17% (per site) (per Mys)
- Substitution model (p): JC69
- Evolutionary time to simulate (T): 200 MYA
- Length of initial genomic sequence ($seq.len$): 10,000 bp
- Transposition rate ($Tr.rate$): 10 mutations per insertion
- Threshold to deactivate a TE (PID): 90%

An assumption that any site in the sequence has the same neutral mutation rate of 0.17% per site per million years [13] is made for simplicity. Similarly, the same substitution model is applied to each site in the nucleotide sequence for simplicity as well. In our simulation, the JC69 model [14] (one of the most common DNA substitution models) is applied to the genomic sequence. As mentioned in the study of [15], no Alu elements with more than 10% mutations were active in the cell culture in [16]. Therefore, the threshold of percent identity to deactivate a TE in the simulation is set to be $PID = 90\%$. The transposition rate, denoted as $Tr.rate$, for the Alu elements has been estimated as approximately 1 insertion for every 20 births in humans [17]. We assume that transposons have the same mutation rate as their surrounding DNA loci after they are inserted into the genome. We also assume that both the transposition rates and mutation rate are constant over time. Given a neutral mutation rate of 160 mutations per diploid genome per generation in human [18] (this is is comparable to the mutation rate of 0.17% per site per Mys in [13]), the transposition rates of Alu elements in the human genome can then be converted and represented relative to the sequence evolution in our simulation as:

$$\begin{aligned} Tr.rate &= \frac{\text{160 mutations (per diploid genome) (per generation)}}{\text{1/20 } Alu \text{ insertion (per diploid genome) (per generation)}} \\ &= 3,200 \text{ mutations (per } Alu \text{ insertion).} \end{aligned}$$

However, to make the time taken to execute the simulation practical, only a small set of 20 TEs (2% of the TEs in the human genome) will be simulated on a small genome of length 10,000 bp ($3.33 \times 10^{-4}\%$ of the human genome size). Moreover, in order to generate a large number of insertions and interruptions in a practical amount of time in the simulation, the transposition rate is set to $Tr.rate = 10$ mutations (per insertion) for simplicity.

It should be noted that this simulation only imitates the transposition of retrotransposons (transpose using copy-and-paste mechanism), not DNA transposons (transpose using cut-and-paste mechanism). This is because each insertion of retrotransposons is stable through evolutionary time, and is a "fossil" of a unique transposition event. The transposition of DNA transposons involves not only insertions, but also excisions, with different rates, which requires more studies.

C. Inputs

The simulation takes TEs and their properties as input, including TE consensus sequences, ages, and harmful regions of these TEs.

i. TE consensus sequences: a list of n TEs whose propagation will be simulated along with their consensus sequences. The consensus is randomly generated with a length of 30 bp (10% of the length of Alu).
ii. TE ages: the list of n TEs are input together with their age of activities (ranging from 200 Mys to 30 Mys). The TEs will be activated once the current time reaches their age (when the TEs start appearing in the genome).
iii. Harmful regions: certain genomic positions in the consensus sequences of the TEs are called harmful regions. If mutations occur within the harmful regions of a TE, the activity fraction of this TE will be decreased in the simulation (mutations in certain regions of TEs are more likely to affect activity [15]). According to the harmful regions of real $AluY$ elements calculated in [15], the harmful regions covered 34.5% of the $AluY$ consensus sequence. Therefore, in the simulation, 30% of the consensus sequences are marked as harmful regions, where the positions of the regions are randomly generated.

D. General steps of the simulation

As previously discussed, the TE transposition simulation is based upon the sequence evolution simulation, where random mutations are introduced into the sequence for each time step iteratively. For every $Tr.rate$ (a transposition rate described in terms of the number of mutations) mutations, introduce a TE insertion. The inserted TE is replicated from either a newly activated TE from the TE database (when the current time matches the age of that TE) or from a randomly selected active TE copy (from its activity fraction) that already existed in the genome. Each active TE existing in the genome has an attribute called activity fraction, denoted as $activeFrac$, which is dynamically calculated by the current percent identity, the number of mutations that occurred within the harmful regions, and the lifespan of that TE. The mutations and insertions are repeated until the simulation time is exhausted.

E. Simulation results

Although the simulation will be run 10 times and aggregate statistics will be collected, one simulation will be described in detail first as an example. This allows for a more detailed discussion.

A simulation of 20 TEs was run for $T = 200$ Mys using the parameters in Section III-B. The consensus and harmful regions (that cover 30% of the length of the consensus) of these TEs are randomly generated. The activation and deactivation time and the lifespan of each TE in the simulation are in Table I (from oldest to youngest). Note that the TEs are labelled intentionally to be consistent with their age for simplicity; for example, $TE1$ is the oldest, and $TE20$ is the youngest. The example is continued in the next section.

TE name	Input age order (oldest to youngest)	Year of activation (MYA)	Year of deactivation (MYA)	Lifespan (Mys)	# of fragments in genome
TE1	1	199	138	61	80
TE2	2	195	171	23	6
TE3	3	189	168	22	3
TE4	4	185	151	34	1
TE5	5	180	112	68	77
TE6	6	170	119	50	13
TE7	7	160	89	71	58
TE8	8	150	86	64	59
TE9	9	140	69	71	19
TE10	10	130	55	74	122
TE11	11	120	97	22	1
TE12	12	110	62	48	7
TE13	13	100	63	37	9
TE14	14	90	13	76	129
TE15	15	80	39	41	5
TE16	16	70	8	61	22
TE17	17	60	0	60	60
TE18	18	50	0	50	53
TE19	19	40	0	40	33
TE20	20	30	0	30	29

TABLE I

THE INPUT AGE ORDER, ACTIVATION TIME, DEACTIVATION TIME AND LIFESPANS OF TES IN THE SIMULATION. THE COLUMN OF INPUT AGE ORDER IS MARKED IN RED, WHICH WILL BE COMPARED TO THE PREDICTED AGE ORDERS LATER.

IV. VERIFICATION OF PREDICTED ORDERING BY SIMULATION

An interruption matrix of sequential interruptions is calculated from the simulation, which is then fed into the sequential interruption model to verify the prediction of the model.

First, the interruption matrix (IM) is fed into the tabu search of the LOP to predict an age order from IM. The predicted age order is then compared to the input age order of TEs, and the correlation of the comparison will be reported reflecting the accuracy of the predicted ages against the input known ages.

The predicted age order from the interruption matrix is calculated and shown in Table II. Note that TE4 and TE11 were not involved in any interruptions, so their relative ages are not predicted.

TE name	Predicted age order calculated by tabu search from IM (oldest to youngest)
TE1	3
TE2	5
TE3	1
TE5	2
TE6	6
TE7	7
TE8	4
TE9	9
TE10	8
TE12	12
TE13	13
TE14	10
TE15	11
TE16	14
TE17	15
TE18	18
TE19	17
TE20	16

TABLE II

THE RELATIVE AGE ORDER CALCULATED BY TABU SEARCH FROM THE INTERRUPTION MATRIX OF THE SEQUENTIAL INTERRUPTION MODEL. NOTE THAT TE4 AND TE11 WERE NOT INVOLVED IN ANY INTERRUPTIONS, SO THEIR RELATIVE AGES ARE NOT PREDICTED.

To quantify how much the predicted order is correlated with the input order, the Pearson's coefficient of correlation is calculated between the predicted age order and the input TE age order as:

$$\rho = 0.9401445,$$

which indicates that the predicted order and the input order have strong positive correlation. Furthermore, Fig. 2 shows the comparison between the predicted age order and the input TE age order. As previously mentioned, it is reasonable that the two orders in comparison are distributed "around" the diagonal line, as TEs have overlapped lifespans. It can be seen that the predicted age order calculated by tabu search agrees well with the input age order.

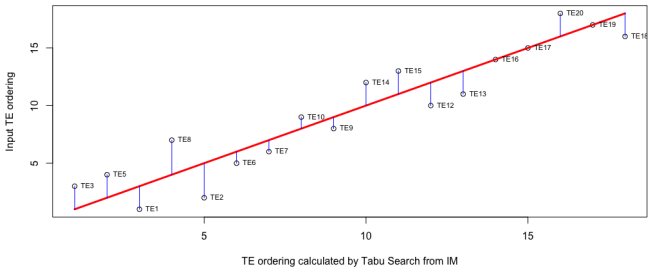

Fig. 2. A comparison between the input age order and the predicted age order calculated by tabu search from the simulated interruption matrix. The x-axis is the predicted relative age order, the y-axis is the input relative age order, and the diagonal line marked in red represents all points that agree between the predicted and actual order.

To further evaluate how well the sequential interruption model predicts the relative ages of TEs in general, the simulation was run 10 times with the same parameters for 20 TEs for $T = 200$ Mys using the same data input (TE consensus, age order, and harmful regions). The same workflow is applied to all the simulated data to compare the predicted orders from the sequential interruption model to the input order. The correlations are listed in Table III.

	ρ
simulation 1	0.936
simulation 2	0.889
simulation 3	0.686
simulation 4	0.611
simulation 5	0.641
simulation 6	0.734
simulation 7	0.719
simulation 8	0.787
simulation 9	0.692
simulation 10	0.681
average	0.738

TABLE III

CORRELATIONS CALCULATED FROM TEN SIMULATIONS.

The correlations in the table suggest that the predicted age order from the sequential interruption model agree well with the input order with an average correlation of 0.738.

V. DISCUSSION

We have previously created the TE sequential interruption model in [7] based on the abundance of TEs interrupting other

TEs, and the problem of predicting TE age was formulated by this model as a well-studied matrix problem — the linear ordering problem. In this paper, tabu search, one of the most efficient known meta-heuristic methods for LOP, was used to solve the problem very efficiently. As discussed, though [5] did not report how long it took the repositioning method in their interruptional analysis to solve the problem on the reduced matrix (of size 405), it is likely long, as only a portion of TEs were solved. In contrast, the tabu search solves the LOP on the full size matrix (of size 1,015) in just 38 seconds, while achieving better results when restricting to the elements common in both methods. This efficiency should allow the method to be usable for all TEs in every genome. Overall, interruptional analysis provides a novel analysis of the evolutionary history of some of the most abundant and ancient repetitive DNA elements in mammalian genomes by analyzing only a single genome, which is important for understanding the dynamic forces that shape the genomes during evolution.

Furthermore, a simulation is developed in the paper, which simulates the evolutionary process of how TEs propagate through time, built on top of the existing `PhyloSim` package that simulates sequence evolution. It is based on several assumptions, such as the mutation rate being constant both through all genomic sites including inserted TEs, as well as through evolution. It also assumes that the transposition rate is constant for all the TEs in the simulation. The simulation can be extended in different ways to include more parameters and controls in order to closer approximate realistic situations. The simulation method is also useful for future TE prediction methods for verification.

VI. Conclusions

In conclusion, the LOP and tabu search in particular as per the sequential interruption model is more practical than existing approaches while achieving better results in predicting the relative ages of TEs. The TE remnants in the simulated genome and their actual ages in the simulation are used to verify the sequential interruption model. As a performance measurement of the prediction model, the predicted relative ages calculated by tabu search based on the sequential interruption model shows a strong positive correlation with the input age order of TEs. With these newly established methods, tabu search can be applied to any genome with TEs. Moreover, the big speed improvements allow the possibility for comparative analysis of TEs in any genome, and even in multiple genomes, in order to advance our understanding about evolution of multiple species where common TEs exist between them.

Acknowledgment

The authors would like to thank Dr. Raymond Spiteri for helpful discussions regarding the linear ordering problem.

References

[1] D. J. Finnegan, "Eukaryotic transposable elements and genome evolution," *Trends in Genetics*, vol. 5, pp. 103–107, 1989.

[2] C. Feschotte and E. Pritham, "DNA transposons and the evolution of eukaryotic genomes," *Annual Review of Genetics*, vol. 41, pp. 331–368, 2007.

[3] J. Jacob-Hirsch, E. Eyal, B. A. Knisbacher, J. Roth, K. Cesarkas, C. Dor, S. Farage-Barhom, V. Kunik, A. J. Simon, M. Gal *et al.*, "Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders," *Cell Research*, vol. 28, pp. 187–203, 2018.

[4] P. A. Larsen, M. W. Lutz, K. E. Hunnicutt, M. Mihovilovic, A. M. Saunders, A. D. Yoder, and A. D. Roses, "The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease," *Alzheimer's & Dementia*, 2017.

[5] J. Giordano, Y. Ge, Y. Gelfand, G. Abrusán, G. Benson, and P. Warburton, "Evolutionary history of mammalian transposons determined by genome-wide defragmentation," *PLoS Computational Biology*, vol. 3, no. 7, p. e137, 2007.

[6] A. V. Zimin, D. Puiu, R. Hall, S. Kingan, B. J. Clavijo, and S. L. Salzberg, "The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum," *Gigascience*, 2017.

[7] L. Jin and I. McQuillan, "Computational modelling of interruptional activities between transposable elements using grammars and the linear ordering problem," *Soft Computing*, vol. 20, no. 1, pp. 19–35, 2016.

[8] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP Completeness*. San Francisco: Freeman San Francisco, CA, 1979, vol. 174.

[9] F. Glover and M. Laguna, *Tabu Search**. New York: Springer, 2013.

[10] J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase update, a database of eukaryotic repetitive elements," *Cytogenetic and Genome Research*, vol. 110, no. 1-4, pp. 462–467, 2005.

[11] B. Sipos, T. Massingham, G. E. Jordan, and N. Goldman, "PhyloSim-Monte Carlo simulation of sequence evolution in the R statistical computing environment," *BMC Bioinformatics*, vol. 12, no. 1, p. 104, 2011.

[12] S. Mahadevan, "Monte carlo simulation," *Mechanical Engineering-New York and Basel-Marcel Dekker-*, pp. 123–146, 1997.

[13] H. Khan, A. Smit, and S. Boissinot, "Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates," *Genome Research*, vol. 16, no. 1, pp. 78–87, 2006.

[14] T. H. Jukes, C. R. Cantor *et al.*, "Evolution of protein molecules," *Mammalian Protein Metabolism*, vol. 3, no. 21, p. 132, 1969.

[15] L. Jin, I. McQuillan, and L. Li, "Computational identification of harmful mutation regions to the activity of transposable elements," *BMC Genomics*, vol. 18, no. 9, p. 862, 2017.

[16] E. A. Bennett, H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, O. Weichenrieder, and S. E. Devine, "Active Alu retrotransposons in the human genome," *Genome Research*, vol. 18, no. 12, pp. 1875–1883, 2008.

[17] R. Cordaux, D. J. Hedges, S. W. Herke, and M. A. Batzer, "Estimating the retrotransposition rate of human Alu elements," *Gene*, vol. 373, pp. 134–137, 2006.

[18] M. W. Nachman and S. L. Crowell, "Estimate of the mutation rate per nucleotide in humans," *Genetics*, vol. 156, no. 1, pp. 297–304, 2000.