Genetic distance between complex repeats

Luca Ferretti[1]*, Aurora Ruiz-Herrera[2], Alice Ledda[3]

[1]The Pirbright Institute, Woking, United Kingdom.
[2] Institut de Biotecnologia i Biomedicina and
Departament de Biologia Cellular, Fisiologia i Immunologia,
Universitat Autònoma de Barcelona, Bellaterra, Spain.
[3]Department for Infectious Disease Epidemiology,
Imperial College London, London, United Kingdom.

Abstract

Complex nucleotide or aminoacid repeats with long units play an important role in proteins. The evolutionary analysis of these variants is challenging due to genetic diversity within repeat units as well as variability in the arrangement of different units along the repeat sequence. Here we present a new approach for the computation of genetic distances between complex repeats. This method takes into account evolutionary processes including point mutations, insertions and deletions of repeat units, as well as duplication of single units. We provide an algorithm for the computation of these distances along with the corresponding global pairwise alignment of repeats. This approach opens the way for new insights into the evolutionary history of polymorphic repeats.

Introduction

Reconstructing phylogenies of repetitive regions has always been a challenge for evolutionary biologists. In the dawn of the genomic era these regions were masked

*Email: luca.ferretti@gmail.com

in reconstructing phylogenies due to the intrinsic difficulties they posed [2]. The evolution of repetitive regions is mostly shaped by peculiar genomic processes like non-homologous recombination, replication slippage etc [3]. These processes often play a major role in the evolution of such regions compared to point mutations. This increases the difficulties in modelling the evolution of such genomic regions, as these processes need to be included in realistic approaches [1].

Existing models are focused on specific types of repeats such as microsatellites, Copy Number Variants etc, where repeat units are similar or identical and the only degree of freedom is their length in terms of the number of repeated units. A more challenging and richer family of repeats is represented by complex satellite repeats with long repeat units (tens of bp or aminoacids) with some degree of internal variability between similar units, as well as variability in the repeat composition in terms of groups of similar units. Only a few of the existing bioinformatic tools are able to deal effectively with this kind of repeats, and most of them provide alignment of protein repeats [6, 7, 5]. To our knowledge, the only practical approach with a currently available implementation is repeat-aware Multiple Sequence Alignment of such repeats with ProGraphMSA+TR [10].

On long-term scales, all units of a repeat often derive from a single ancestral unit via point mutations and duplication/slippage. This is true even for complex repeats with different types of units [1]. However, on short evolutionary scales (between close species or within species), insertions and deletions of highly diverged units represent clearly different processes with respect to recent duplication/slippage followed by divergence via point mutations. Existing methods for repeat alignment do not discriminate between these processes, and the difference between these processes is not taken into account in the computation of the genetic distance.

Here we develop a new approach for the computation of genetic distances for complex repeats. Our approach includes point mutations, insertions and deletions of whole repeat units, as well as duplications of single repeat units as a consequence of non-homologous recombination, slippage or related biological processes. Single-unit duplications and indels can have different weights and therefore contribute differently to the genetic distance. We also present a version of the Needleman-Wunsch algorithm [4] to compute the genetic distance between pairs of repeats, as well as their optimal pairwise global alignment according to the new genetic distance defined here. The result of this algorithm is a modified alignment that allows for further evolutionary analyses on single-unit duplication/slippage.

Methods

In our approach, repeat units are treated as fundamental blocks. Only processes preserving the integrity of each block are considered. The identification of repeat units is outside the scope of this paper. The choice of the repeat units is left to the user.

We assume that all repeat units share a high sequence similarity. Hence, Multiple Sequence Alignment of all units from all repeats can be easily obtained from any alignment method. It is therefore straightforward to compute pairwise genetic distances $\mu(u, u')$ between units u and u'. Many possible definitions of genetic distance between biological sequences exist; the choice of the most appropriate one is left to the user.

Definition of distance

We define the genetic distance between repeats as an edit distance, i.e. as the minimum cost to change a repeat to another through a series of elementary operations. The cost is defined as the sum of the weights of all elementary steps. In our case, elementary operations are inspired by biological processes. They are:

- point mutations, small indels and other within-unit processes (weight w_m);
- insertions and deletions of whole units (weight w_i);
- single-unit duplication/slippage (weight w_s).

The edit distance is minimised by one or possibly several alignments between repeats $r = (r_1, r_2, r_3 \dots)$ and $r' = (r'_1, r'_2, r_3 \dots)$. For each of these alignments, we denote by $\mathcal{M}$ the set of all matching units between r and r' (the uth unit in the repeat r corresponds to the $m(u)$th unit in r'). $\mathcal{I}$ and $\mathcal{I}'$ denote the sets of inserted units in r and r' respectively. Finally, $\mathcal{S}$ denotes the set of duplicated units in r, where the uth unit is the result of a duplication of the $s(u)$th unit (which could be either the $(u+1)$th or the $(u-1)$th unit). $\mathcal{S}'$ and $s'(u)$ denote the same quantities for r'.

Our definition of the distance between r and r' is

$$d(r,r') = \sum_{u\in\mathcal{M}} w_m\mu(r_u, r'_{m(u)}) + \sum_{u\in\mathcal{I}} w_i + \sum_{u\in\mathcal{I}'} w_i + \sum_{u\in\mathcal{S}} [w_s + w_m\mu(r_u, r_{s(u)})] + \sum_{u\in\mathcal{S}'} [w_s + w_m\mu(r'_u, r'_{s'(u)})] \tag{1}$$

This distance can either be derived from multiple repeat alignment [10] or it can be directly computed using the algorithm discussed in the next section.

Needleman-Wunsch algorithm with single-unit duplications

We present a modification of the classical Needleman-Wunsch algorithm [4] to include single-unit duplications in the computation of the distance. For each pair of repeats, this algorithm provides their genetic distance as well as the corresponding global alignment.

Our algorithm mirrors closely the standard Needleman-Wunsch algorithm. The differences with respect to the standard version of the algorithm lie in the choice of weights and in the way insertions/deletions are handled.

Our selection of weights is as follows:

- mismatch between units r_u and $r'_{u'}$: $w_m\mu(r_u, r'_{u'})$
- insertion of a whole unit r_u in repeat r:

$$\min(w_i, w_s + w_m\mu(r_u, r_{u-1}), w_s + w_m\mu(r_u, r_{u+1}))$$

Given these weights, the distance is then computed as in the standard algorithm.

Furthermore, when the best partial alignment results from an insertion in either of the two repeats, we consider the minimum among the values of w_i,

$w_s + w_m \mu(r_u, r_{u-1})$ and $w_s + w_m \mu(r_u, r_{u+1})$. These values correspond to three different scenarios:

Value	Process	Gap-filling symbol
w_i	standard insertion	$-$
$w_s + w_m \mu(r_u, r_{u-1})$	single-left-unit duplication	$<$
$w_s + w_m \mu(r_u, r_{u+1})$	single-left-unit duplication	$>$

The alignment is computed as in the standard algorithm, but for each insertion, the corresponding gap-filling symbol from the above table is then used to represent the gap in the alignment. In this way, it is possible to separate single-unit duplications from actual insertions/deletions. Note that there could be ties among the above values; in this case, multiple optimal alignments are possible.

Discussion

In this paper we present a simple approach to the computation of genetic distances for complex repeats. Phylogenetic reconstruction of their evolutionary history can then be extracted using Neighbour-Joining or other distance-based methods [9]. A byproduct of this approach is an Needleman-Wunsch algorithm for global alignment of repeats. This algorithm can be easily extended to a version of the Smith-Waterman algorithm [8] to provide local alignment of highly divergent repeats.

The key feature of our approach is that we consider single-unit duplications in addition to insertions/deletions. This implies that the method depends on two parameters (w_i/w_m for indels and w_s/w_m for duplications). More complicate processes could be included, such as duplication from non-adjacent units or large insertions/deletions, at a price of extra parameters to be tuned. A solution could be to add extra features such as separate "gap opening" and "gap extension" penalties instead of a single insertion cost, and similar parameters for duplications. This could improve the handling of large non-homologous recombination events.

A structural limitation of our approach is that complex rearrangements would be captured as alignments with many insertions/deletions, implying large genetic distances. Another limitations is that by treating units as irreducible blocks, we neglect processes involving only subparts of units or shifted units. Processes that affect only a part of the unit sequence, such as within-unit recombination, shifted insertion/deletion/duplication/slippage and so on, are therefore approximated as occurring at one of the extremes of the unit or as involving multiple units. Despite these drawbacks, our method provides a fast and effective estimate of genetic distances for polymorphic repeats from related species or populations.

Acknowledgments

We thank Philipp Schiffer, Covadonga Vara and two anonymous referees of Molecular Ecology for comments and suggestions. A R implementation of the algorithms described here is provided at `https://github.com/lucaferretti/RepeatDistance`.

References

[1] Maria Anisimova, Julija Pečerska, and Elke Schaper. Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. *Frontiers in bioengineering and biotechnology*, 3:31, 2015.

[2] Nansheng Chen. Using repeatmasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, 5(1):4–10, 2004.

[3] Philip J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551, 2009.

[4] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[5] Tu Minh Phuong, Chuong B Do, Robert C Edgar, and Serafim Batzoglou. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic acids research*, 34(20):5932–5942, 2006.

[6] Benjamin Raphael, Degui Zhi, Haixu Tang, and Pavel Pevzner. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research*, 14(11):2336–2346, 2004.

[7] Michael Sammeth and Jaap Heringa. Global multiple-sequence alignment with repeats. *Proteins: Structure, Function, and Bioinformatics*, 64(1):263–274, 2006.

[8] Temple F Smith and Michael S Waterman. Comparison of biosequences. *Advances in applied mathematics*, 2(4):482–489, 1981.

[9] Mike Steel. *Phylogeny: discrete and random processes in evolution.* SIAM, 2016.

[10] Adam M Szalkowski and Maria Anisimova. Graph-based modeling of tandem repeats improves global multiple sequence alignment. *Nucleic acids research*, 41(17):e162–e162, 2013.