The Transformation Distance

Jean-Stéphane Varré [1] **Jean-Paul Delahaye** [1] **Éric Rivals** [2]
varre@lifl.fr delahaye@lifl.fr E.Rivals@dkfz-heidelberg.de

[1] Université des Sciences et Technologies de Lille,
Laboratoire d'Informatique Fondamentale de Lille - URA CNRS 369
UFR IEEA - Bât M3, 59655 Villeneuve d'Ascq Cedex, France

[2] Theoretische BioInformatik, Deutsches Krebsforschungzentrum (DKFZ),
Im Neuenheimer Feld 280, Heidelberg 69120, Germany

Evolution operates molecular alterations of two types: *punctual* mutations, namely insertion, deletion or substitution of one residue, and *segment-based* modifications: duplication, displacement, deletion or insertion of a segment of the molecule. Up to now, distance or dissimilarity measures used to compare sequences consider only punctual mutations. To our knowledge, no measure attempts to quantify dissimilarity by assessing segment-based differences, nor by describing the differences between two sequences with an edit-script containing such segment operations. Consequently, sequence comparison is often performed on similar parts of the sequences, like structurally or functionally related domains of proteins. Even if they correspond to complete biological entities like whole gene or protein, entire sequences are not compared, or only in case of high similarity. With such restrictions, one forgets some evolutionary meaningful informations written in the molecules.

We propose a new similarity measure which considers four types of segment-based operations: duplication (i.e., copy and re-insertion at another or at the same location), displacement, deletion or insertion of a segment. Only exact segment are subject to duplication. Given a source sequence A and a target sequence B, there is always at least one list of operations which transforms sequence A into sequence B, but usually there are many. Figure 1 displays a transformation. Even when A and B have no common segment, a list containing the insertion of the entire sequence B achieves the transformation in one operation. A list of operation is called *a transformation* and our measure *the transformation distance.*

Here, a measure that would simply count the number of operations in the list is unsuitable. The weight of an operation should depend on the length of the segment on which it is applied. We use *the description length* of an operation as its weight, then the distance between A and B is *the minimum description length* among all possible lists of operations transforming A into B. The coding used to *describe* an operation is carefully chosen to include only necessary informations. A displacement is coded by a triplet of integers: the segment's positions in A and B, and its length. Given A and B, this distance approximates the quantity of information necessary to reconstruct B from A. The idea of comparing description length of transformations comes from the following property of the Algorithmic Information Theory [4] (which was already used in biological applications [6]): *shortest descriptions are the more probable explanation of the phenomenon.* In our case, the phenomenon is the evolution from A to B thanks to segment-based events.

As a segment can be one character long, the transformation distance copes with cases where punctual distances are usually applied. It is thus more complex to compute, but also more general. Our distance cannot be directly compare to reversal distances [2] because a different operation, the segment reversal, is considered. A main difference is that the transformation distance identifies the segments, while reversal distances require known genes (i.e. peculiar segments) as input.

Algorithm

We designed and implement a heuristic algorithm to compute the transformation distance between DNA sequences (it is available on www[1].) First, it computes the subset of transformations which maximize the length of the segments involved in copies or displacement (common segments of A and

[1] http://www.lifl.fr/~varre/phylogeny.html

B). The size of those is required to be greater than the average size of random common segments (i.e., maximal common segment between two random sequences): $log_4|s|$ where $|s|$ is the sum of the sequences length. We known from the Algorithmic Information Theory that random common segments cannot belong to a minimum description length transformation. In a second step, the algorithm searches for the transformation of Minimum Description Length (MDL), which is its result.

Biological applications

A natural application is the phylogeny of species. Given of set of n sequences, we compute the matrix of pairwise transformation distances which serves for building the phylogeny of the species. We investigate the phylogeny of Cetacean with the data proposed in the work of Arnason and Gullberg [1]. The comparison is done with the sequences of cytochrome b which show no specific re-organisation of the molecules: a usual case for punctual distances. We build the phylogenetic tree using the Fitch's method of tree construction and recover the classical phylogeny of Cetacean (data not shown)[2].

We also study the sequence variability in tobacco retrotransposons. The data, taken from [5], contains twenty samples of Tnt1 retrotransposon for each of the seven tobacco species. We also compute the matrix of pairwise distances to elude the relationships between all retrotransposons (data not shown). It suggests that the great variability of a specific region is due to a lot of segment duplications and displacements.

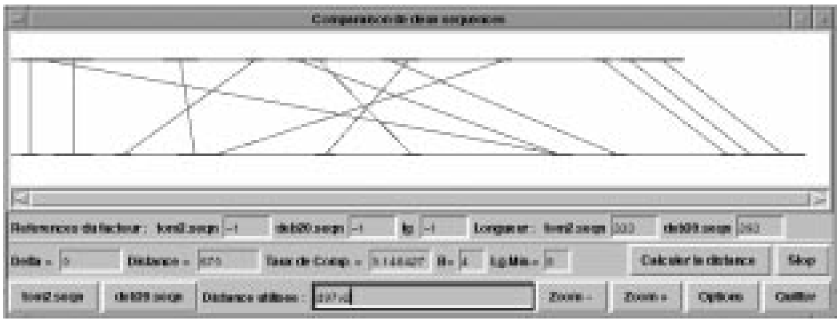

Figure 1: Comparison between retrotransposons of *nicotania tomentosiformis* (top) and *nicotania debneyi* (bottom). A line joins each (dark grey) segment in common to A and B. The others segments (light grey), are either deleted (those in A and not in B) or inserted (those in B but not in A.)

On figure 1, we display the result for a pairwise comparison of two species. It shows the minimal transformation between two sequences, say A and B. The transformation distance from A to B is small because they share a lot of duplicated or displaced segments. The same comparison with punctual distances finds the sequences very distant. The transformation distance is a more adequate measure to study sequences relationships when segment re-organisation is suspected.

References

[1] Ùlfur Àrnason and Anette Gullberg, "Relationship of baleen whales established by cytochrome b gene sequence comparison", *Nature*, 367:726-728, February 1994.

[2] V. Bafna and P. Pevzner, "Genome rearrangements and sorting by reversals", Proc. 34th FOCS, IEEE, 148-157, 1993.

[3] Wen-Hsiung Li and Dan Graur, *Fundamentals of Molecular Evolution*, Sinauer Associates Inc., 1991.

[4] M.Li and P.M.B.Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Spinger-Verlag, New-York, 2nd Edition, 1997.

[5] Josep M.Casacuberta, Samantha Vernhettes and Marie-Anghle Grandbastien, "Sequence variability within the tobacco retrotransposon Tnt1 population", *EMBO Journal*, 14(11):2670-2678, 1995.

[6] Éric Rivals, Max Dauchet, Jean-Paul Delahaye and Olivier Delgrange, "Compression and genetic sequences analysis", Biochimie vol. 78, 1996.

[2] The phylogeny of Arnason and Gullberg has some slight differences. The major difference is that sperm whale is placed before separation between dolphins and others Cetacean.