Résumé

We describe here a novel approach to the reconstruction of the duplication history of tandemly repeated genes, based on a model of duplication by unequal recombination. Starting with a description of our model, we follow by presenting duplication histories and duplication trees, as mathematical objects describing the evolution of tandemly repeated genes. We then show how, given a set of ordered nucleotide sequences, it is possible to reconstruct the duplication trees explaining these sequences, so as to optimize a parcimony criterion. Studies on immunoglobulins et T-cell receptors data sets clearly show the validity of our model of duplication and the relevance of our approach.

Classical phylogenetic analysis studies the relationships between species based on the comparison of a single gene. Its main goal is to reconstruct a tree which represents the history of speciations. The problem we describe here is different : we aim at reconstructing the duplication history of a single gene within a single genome and we uniquely consider long (several kilobases) and tandemly arranged sequences, where each one contains a single gene. Assuming our sequences were not affected by gene conversion events, and our loci did not undergo any deletions, we introduce a simple model of duplication based solely on unequal recombination (unequal recombination is commonly acknowledged as the primary mechanism responsible for tandem duplications). Our model of duplication allows simple duplications (a gene is duplicated and inserted adjacent to the initial gene), and bloc duplications (a bloc of 2 or n sequences is duplicated, and inserted near the initial bloc). Although identical just after duplication, these sequences diverge over time as they accumulate their own mutations. We then define three types of mathematical objects to describe the evolution of these clusters of tandemly repeated genes. First, we define what we call a time-valued duplication history, i.e. a description of the real duplication history. Since we cannot generally rely on the moleculer clock hypothesis, inferring a time-valued duplication history from nucleotide sequences is not possible. In particular, the position of the root and the order in which duplications occurred cannot be recovered from DNA sequences. Consequently, we can only reconstruct what we call a duplication tree, i.e an unrooted phylogeny whose topology is compatible with at least one duplication history. According to the model of duplication, the root of a duplication tree can only be situated somewhere (but not everywhere) in the tree between the most distant repeats on the locus. When rooted at one of its allowed branches, a duplication can be transformed into what we call an ordinal duplication history, i.e. a history in which duplication events are partially ordered. Although a duplication tree is a phylogeny, it is easy to show that not all phylogenies can be duplication trees. We use an algorithm we called PDT (for PossibleDuplicationTree) to determine whether a given phylogeny with ordered leaves can be a duplication tree or not. This algorithm provides us with a mathematical characterisation of duplication histories and duplication trees. We also use the PDT algorithm to show that, for a given number of tandemly repeated sequences, the number of duplication trees is largely inferior to the number of distinct phylogenies. Given this model of duplication, we use an exhaustive search procedure to reconstruct duplication trees : given a set of nucleotide sequences, we compute the parcimony value of every possible duplication tree, and we select those which minimize this value. To speed up the reconstruction (especially when dealing with large numbers of repeated genes), we use a faster (but not guaranteed to find the optimal tree) search procedure based on a greedy heuristics : starting with a tree made from the first three repeats, our procedure iteratively inserts new repeats onto the growing tree, such that each resulting tree minimizes the parcimony value. The procedure stops when all repeats are inserted. We applied this model and these search procedures to two human loci containing tamdemly repeated immunoglobulins and T-cell receptors genes : the IGLC and TRGV loci. We showed for both these loci that the duplication tree found by our exhaustive search procedures corresponds

to the most parcimonious phylogeny. Since the probability of a phylogeny being a duplication tree is small (0.04 in the TRGV case), this constitutes a strong validation of our initial hypothesis concerning the duplication mechanisms. Besides, the heuristics-based search reconstructs the same duplication tree as the exhaustive search, but in a much faster way. These results keep stable to a bootstrap analysis, indicating that this identity between the most parcimonious duplication tree and the most parcimonious phylogeny is not fortuitous. Compatibility of our reconstructed trees with known polymorphisms (two genes are missing in some individuals) in the TRGV locus provides further evidence that our reconstruction can provide good insights into the duplication histories of tandemly repeated genes.