

A variant of the tandem duplication - random loss model of genome rearrangement

Mathilde Bouvel and Dominique Rossin

March 29, 2007

In the usual models for genome rearrangement, duplications and losses of genes are not taken into account. There were attempts to incorporate them to the classical models, but the consecutive combinatorial complexity of the models so obtained made their study quite difficult. Following [1], we focus on the duplication-loss problem by considering the *tandem duplication - random loss model* for genome rearrangement in which genomes are modified *only* by duplications and losses of genes.

One *step* of tandem duplication - random loss, or duplication-loss for short, consists in (1) the tandem duplication of a contiguous fragment of the genome, *i.e.*, the duplicated fragment is inserted immediately after the original fragment, and (2) the loss of one of the two copies of every duplicated gene. We assume that the loss occurs immediately after the duplication of genes, which is, on an evolutionary time-scale, a good approximation to reality. The *width* of a step is the number of duplicated genes.

From a formal point of view, a genome consisting of n genes is modelled by a permutation $\pi \in S_n$ of the set of integers $\{1, 2, \dots, n\}$. In [1], they define the cost of a duplication-loss step of width k to be α^k , $\alpha \geq 1$ being a constant parameter. They suggest that other cost functions can be considered, and in particular affine functions. Here, we consider a *piecewise constant* cost function: the cost of a step of width k is 1 if $k \leq K$ and is infinite for $k > K$, for some fixed parameter $K \in \mathbb{N} \cup \{\infty\}$. Both models are generalizations of the *whole genome duplication - random loss model*: it corresponds to the case $\alpha = 1$ in the model of [1], $K = \infty$ in our model.

In the model described above, we obtain results of two kinds.

First, we consider the class of permutations obtained from $12\dots n$ (for any n) after p steps of width at most K , for some constant parameters p and K . This class is denoted $\mathcal{C}(p, K)$. We show that $\mathcal{C}(p, K)$ is a class of pattern-avoiding permutations. In the case $p = 1$, we give a precise description of the basis B of excluded patterns: $B = \{321, 3142, 2143\} \cup D$, D being the set of all permutations of S_{K+1} that do not start with 1 nor end with $K+1$, and containing exactly one descent. In particular, B is of cardinality $3 + 2^{K-1}$ and contains patterns of size at most $K+1$. For the general case, we show that $\mathcal{C}(p, K)$ is a class of pattern-avoiding permutations whose basis contains patterns of size at most $(2kp + 3)(k + 1)$.

A second point of view is to examine how many steps of a given width are necessary to obtain any permutation of S_n starting from $12\dots n$. Namely we fix a width K and a length n and search for the number p such that any permutation of S_n can be obtained from $12\dots n$ in at most p duplication-loss steps of width at most K . We describe an algorithm computing a possible scenario of duplications and losses for any $\pi \in S_n$, this scenario involving $\mathcal{O}(\frac{n}{K} \log K + \frac{n^2}{K^2})$ duplication-loss steps in the worst case and on average. We also show that $\Omega(\frac{n}{K} \log K + \frac{n^2}{K^2})$ steps are necessary (in the worst case and on average) to obtain any permutation of S_n from $12\dots n$.

References

- [1] K. Chaudhuri, K. Chen, R. Mihaescu, and S. Rao. On the tandem duplication-random loss model of genome rearrangement. SODA, pages 564 – 570, 2006.