Chapter 5

Summary

Mutations play a decisive role among the fundamental mechanisms of molecular evolution as they provide the raw material for the emergence of genetic variation. In contrast to nucleotide substitutions, the nature of DNA insertions and deletions is far less understood. The aim of this thesis was to elucidate the characteristics, origin, and consequences of short DNA insertions and deletions in genome evolution.

We utilized multiple alignments of human, chimp, and rhesus for a genome-wide identification of short insertions and deletions that have recently occurred in the human lineage since speciation from the common ancestor with chimp. We showed that the majority of insertions are tandem duplications of directly adjacent sequence segments. While this would be expected for events in microsatellites, our striking observation was that tandem duplications are prevalent among all identified insertions, irrespective of whether they occurred in repetitive regions of the genome, or not. In fact, indels in microsatellites were found to comprise only a small fraction of all indels.

Implications of this finding are manifold. For example, we could show that the sequence characteristics of indels and their genomic vicinity often differ from the signatures expected for unequal crossing over or replication slippage – the two processes that are commonly regarded as the predominant mechanisms of indel generation. This led us to the hypothesis that a different mechanism, the nonhomologous end joining after DNA double-strand breaks, possibly also plays a major role in this context.

Tandem duplication insertions might provide an explanation for a distinct statistical feature present in most eukaryotic genomes, so-called long-range correlations in nucleotide composition. Despite their ubiquity, the origin of genomic long-range correlations has been subject of constant debate for more than a decade. Using methods from the fields of stochastic processes and non-linear dynamical systems, we could analytically show that models of local sequence evolution which incorporate tandem duplications belong to a universality class of one-dimensional expansion-randomization systems with generic stationary long-range correlations.

The characteristic exponent that defines the scaling properties of correlation functions was explicitly calculated for several evolution dynamics of the universality class. It is determined by only two effective rates of the particular dynamical model. We also investigated more complex evolutionary scenarios, where rates of the processes vary

in time and along the sequence. We concluded from this analysis that the observed long-range correlations in DNA can indeed result from the repeated action of tandem duplication processes in genome evolution.

The prevalence of tandem duplications among DNA insertion events has profound consequences on the statistics of bioinformatics sequence analysis tools, which often rely on DNA null models to assign significance values to their predictions. We could show that replacing the standard iid null model by one which includes long-range correlations comparable to those observed in genomic sequences substantially changes the p-values of sequence alignment similarity scores.

The concept of duplication-driven evolution is well known for genes. It is perhaps therefore not surprising that tandem duplications are also one of the major modes of mutation on smaller length scales. In fact, we found that in protein-coding regions of the human genome non-frameshifting tandem duplications are presumably less deleterious compared to non-synonymous substitutions. As is already well established for genes and large-scale segmental duplications, tandem duplications might also constitute an important process for the rapid generation of new genetic material and function on smaller scales.