Molecular Morphology - Phylogenetically Informative Characters Derived from Sequence Data
Dissertation
Alexander Donath

A fundamental problem in biology is the reconstruction of the relatedness of all (extant) species. Traditionally, systematists employ visually recognizable characters of organisms for classification and evolutionary analysis. Recent developments in molecular and computational biology, however, lead to a whole different perspective on how to address the problem of inferring relatedness. The discovery of molecules, carrying genetic information, and the comparison of their primary structure has, in a rather short period of time, revolutionized our understanding of the phylogenetic relationship of many organisms. These novel approaches, however, turned out to bear similar problems as previous techniques. Moreover, they created new ones. Hence, taxonomists came to realize that even with this new type of data not all problematic relationships could be unambiguously resolved. The search for complementary approaches has led to the utilization of rare genomic changes and other characters which are largely independent from the primary structure of the underlying sequence(s). These "higher order" characters are thought to be evolutionary conserved in certain lineages and largely unaffected by primary sequence data-based problems, allowing for a better resolution of the Tree of Life.

The central aim of this thesis is the utilization of molecular characters of higher order in connection with their consistent and comparable extraction from a given data set. Two novel methods are presented that allow such an inference. This is complemented with the search for and analysis of known and novel molecular characteristics to study the relationships among Metazoa, both intra- as well as interspecific.

The first method tackles a common problem in phylogenetic analyses: the inference of reliable data set. As part of this thesis a pipeline was created for the automated annotation of metazoan mitochondrial genomes. Data thus obtained constitutes a reliable and standardized starting point for all downstream analyses, e.g. genome rearrangement studies.

The second method utilizes a subclass of gaps, namely those which define an approximate split of a given data set. The definition and inference of such split-inducing indels (splids) is based on two basic principles. First, indels at the same position, i.e. sharing the same end points in two sequences, are likely homologous. Second, independent single-residue insertions and deletions tend to occur more frequently than multi-residue indels. It is shown that trees based on splids recover most of the undisputed monophyletic groups while influence of the underlying alignment algorithm is relatively small.

Mitochondrial markers are a valuable tool for the understanding of small and large scale population structure. The non-coding control region of mitochondrial DNA (mtDNA) often contains a higher amount of variability compared to genes encoding proteins and non-coding RNAs. A case study on a small scale population structure investigates the control region of the European Fire-bellied Toad in order to find highly variable parts which are of potential importance to develop informative genetic markers. A particular focus is placed on the investigation of the evolutionary dynamics of the repetitive region at an inter- and intraspecific level. This includes understanding mechanisms underlying its evolution, i.e. by exploring the impact of secondary structure on slipped strand mispairing during mtDNA replication.

The 7SK RNA is a key player in the regulation of polymerase II (Pol-II) transcription, interacting with at least three known proteins: It mediates the inhibition of the Positive Transcription Elongation Factor b (P-TEFb) by the HEXIM1/2 proteins, thereby repressing transcript elongation by Pol-II. A highly specific interaction with LARP7 (La-Related Protein 7), on the other hand, regulates its stability. 7SK RNA is capped at its 5' end by a highly specific methyltransferase MePCE (Methylphosphate Capping Enzyme). Employing sequence and structure similarity it is shown that the 7SK RNA as well as its protein binding partners have a much earlier evolutionary origin than previously expected. Furthermore, this study presents a good illustration of the pitfalls of using markers of higher order for phylogenetic inference.