Gene order rearrangement methods for the reconstruction of phylogeny

The study of phylogeny, i.e. the evolutionary history of species, is a central problem in biology and a key for understanding characteristics of contemporary species. Many issues in this area can be formulated as combinatorial optimisation problems. This opportunity makes it particularly interesting for computer scientists. A certain property of genetic information gained much interest for reconstruction of phylogeny in recent time: the organisation of the genomes of species, i.e. the arrangement of genes on chromosomes. This kind of data is promising for the study of deep evolutionary relationships because gene arrangements are assumed to evolve slowly. This seems to be the case especially for *metazoan* mitochondrial genomes which are available for a wide range of species.

Formally, gene arrangements are represented as signed permutations, i.e. permutations where each element has an additional sign indicating the orientation of the corresponding gene. At least four types of evolutionary rearrangement events have to be assumed for studying gene order evolution. Inversions reverse a part of the gene order, transpositions move a part of the gene order to a distant position, and inverse transpositions additionally inverse the transposed segment. The fourth rearrangement is the tandem duplication random loss (TDRL) operation. This rearrangement tandem duplicates a continuous part of the gene order followed by the random loss of one copy of each redundant gene.

Two fundamental genome rearrangement problems are studied in this thesis. Both are considered with modified rearrangement models in order to obtain more plausible solutions. It is shown that increased plausibility can be accompanied by an efficient solution.

Two types of modified rearrangement models are explored. The first is motivated by the observation that certain groups of genes are preserved during evolution. Accordingly, these gene groups should be preserved in plausible reconstructions of the course of evolution. In particular the gene groups should be present in the reconstructed putative ancestral gene orders. This can be achieved by restricting the set of rearrangements, which are allowed for the reconstruction, to those which preserve the gene groups of the given gene orders.

The second considered modification of the rearrangement model is extending the set of allowed rearrangement types. Different types of rearrangement operations have shuffled the gene orders during evolution. It should be attempted to use the same set of

rearrangement operations for the reconstruction. Distorted or even wrong phylogenetic conclusions may be obtained otherwise.

The first problem investigated in this work is the inversion median problem (IMP). This is to find a median gene order which can be transformed with a minimum number of inversions into given gene orders. This problem is studied in the modified rearrangement model allowing no inversion in the median scenario to break one of the common intervals in the input permutations, i.e. consecutively appearing gene groups. This is the preserving inversion median problem (pIMP). Three exact algorithms for this problem are presented in this thesis. Algorithms CIP and ECIP are modifications of an existing branch and bound median solver for the IMP. While CIP checks if a generated solution is preserving, ECIP restricts the branch and bound search such that only preserving solutions are generated. The algorithm TCIP for solving the pIMP for k given gene orders is based on the k-signed strong interval tree data structure. This data structure efficiently represents the common intervals of the given gene orders. The relation of the difficulty of the pIMP and the structure of the corresponding k-signed strong interval tree is analysed theoretically. Based on these results, algorithm TCIP is designed such that it can solve a pIMP instance by solving several smaller instances of the IMP.

Properties of the preserving inversion median problem and the three presented algorithms are investigated empirically for simulated and biological data sets. It is shown empirically that common intervals occur often and many common intervals are destroyed by solutions for the IMP, i.e. when common intervals are not considered. It is demonstrated that the algorithms CIP and ECIP can solve the pIMP for most of the simulated and biological data sets in reasonable time. Properties of the corresponding strong interval trees, which are of importance for the run time behaviour of TCIP, are analysed. The empirical study demonstrates that often TCIP has to solve no IMP instances for solving the pIMP. This leads to a linear run time behaviour of TCIP. Note, the preserving inversion median problem is NP-hard. Furthermore, only a few and smaller instances of the IMP have to be solved in the remaining instances. The empirical study clearly shows that TCIP outperforms CIP and ECIP by orders of magnitude. Furthermore, preserving median scenarios can be computed with TCIP even faster than standard (i.e. not necessarily preserving) median scenarios with a state of the art IMP solver. This is remarkable because even computing the minimum number of preserving inversions, necessary to transform one given gene order into another, is an NP-hard problem (whereas the same problem with not necessarily preserving inversions can be solved in polynomial

time). Furthermore, it is demonstrated that algorithm TCIP can be used as a good heuristic for the IMP.

The problem of finding a rearrangement scenario transforming one given gene order into another is studied in the second part of the thesis. The heuristic algorithm `CREx` for computing rearrangement scenarios for pairs of given unichromosomal gene orders is presented. The rearrangement operations inversion, transposition, inverse transposition, and tandem duplication random loss are considered by `CREx`. This covers the biologically evident operations for *metazoan* mitochondrial gene orders. `CREx` reconstructs a rearrangement scenario by identifying patterns in signed strong interval trees which represent the common intervals of gene order pairs. The quality of the `CREx` reconstructions is analysed for simulated data sets generated with several rearrangement models. Parameters of the simulation and properties of the strong interval trees are identified where `CREx` returns results of high quality.

Based on `CREx`, further gene order rearrangement methods for the reconstruction of phylogeny are presented. The algorithm `TreeREx` utilises the pairwise scenarios computed by `CREx` to infer ancestral permutations and genomic rearrangement operations in a given binary phylogenetic tree. `TreeREx` is applied to biological data sets and it is shown that the reconstructed rearrangements are in strong correspondence with published results. Based on the results of `CREx` for simulated data, a new simple method is introduced to explore the rearrangements of a data set without a given phylogenetic tree. The method is applied to the complete data set of *metazoan* mitochondrial genomes. The method obtains very efficiently the complete picture of the rearrangements within the *metazoan* phyla. The returned results are compliant with the literature to a large extent. The presented algorithms `CREx`, `TreeREx`, as well as the new exploration method solved the simulated and biological data sets very efficiently.

Tandem duplication random loss (TDRL) events are important gene order rearrangement operations especially in mitochondrial gene orders. A better understanding of this operation is indispensable for the study of mitochondrial gene orders. Thus, combinatorial properties of the tandem duplication random loss rearrangement are studied in the last chapter of the thesis. The set of all sorting TDRLs and an interesting restricted case of the problem is investigated. Methods for the enumeration of the set of sorting (restricted) TDRLs and closed formulas for calculating the number of sorting (restricted) TDRLs are presented. The results are obtained by an enumeration of binary strings with certain properties. The relevance of the theoretical findings when identifying sequences of TDRLs for real biological data, e.g. mitochondrial gene orders, is shown.