Gene prediction in newly sequenced genomes is a known challenging. Although sophisticated comparative pipelines are available, computationally derived gene models are often less than perfect. This is particularly true when multiple very similar paralogs are present.

The issue is aggravated further when genomes are assembled only at a preliminary draft level to contigs or short scaffolds rather than to chromosomes. However, these genomes deliver valuable information for studying gene families. High accuracy models of protein-coding genes are needed in particular for phylogenetics and for the analysis of gene family histories.

In this dissertation, I established a tool, the ExonMatchSolver-pipeline (EMS-pipeline), that can assist the assembly of genes distributed across multiple fragments (e.g. contigs). The tool in particular tackles the problem of identifying those coding exon groups that belong to the same paralogous genes in a fragmented genome assembly. The EMS-pipeline accommodates a homology search step with a protein input set consisting of several highly similar paralogs as query. The core of the pipeline uses an Integer Linear Programming Implementation to solve the paralog-to-contig assignment problem. An extension to the initial implementation estimates the number of paralogs encoded in the target genome and can handle several paralogs that are situated on the same genomic fragment.

The EMS-pipeline was successfully applied to simulated data, several showcase examples and to deuterostome genomes in a large scale study on the evolution of the arrestin protein family. Especially at high genome fragmentation levels, the tool outperformed a naive assignment method.

Arrestins are key signaling transducers that bind to activated and phosphorylated G protein-coupled receptors and can mediate their endocytosis into the cell. The refined annotations of arrestins resulting from the application of the EMS-pipeline are more complete and accurate in comparison to a conventional database search strategy. With the applied strategy it was possible to map the duplication- and deletion history of arrestin paralogs including tandem duplications, pseudogenizations and the formation of retrogenes in detail.

My results support the emergence of the four arrestin paralogs from a visual and a non-visual proto-arrestin. Surprisingly, the visual ARR3 was lost in the mammalian clades afrotherians and xenarthrans. Segmental duplications in specific clades and the 3R-WGD in the teleost stem lineage, on the other hand, must have given rise to new paralogs that show signatures of diversification in functional elements important for receptor binding and phosphate sensing. The four vertebrate orthology groups show an interesting pattern of divergence of three endocytosis motifs: the minor and major clathrin binding site and the adapter protein-2 (AP-2) binding motif.

Identification of such signatures, of residues that determine specificity between paralogs and are positively selected after duplication was made possible by high quality alignments obtained by genome inquiries, dense species sampling and consideration of fragmented loci from poorly assembled genomes in the framework of the EMS-pipeline, that was established in this dissertation.