High quality gene annotation for deep phylogenetic analysis

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt

von Master of Science Henrike Indrischek

geboren am 6. Mai 1989 in Lutherstadt Wittenberg

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Yves van der Peer (Universiteit Gent, Belgien)
2. Professor Dr. Peter F. Stadler (Universität Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen

der Verteidigung am 12. April 2018 mit dem Gesamtprädikat **magna cum laude**

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

.....

(Ort, Datum)

.....

(Unterschrift)

"Do it with all your energy, you will not fail your aim."

Fortune cookie quote

Acknowledgements

First of all, I would like to thank my scientific supervisors and mentors at the Bioinformatics group in Leipzig – Sonja and Peter – for their ideas and guidance during the project. Thank you Peter for putting so much trust in me to manage and survive in this interdisciplinary field starting without any programming skills! I enjoyed this steady learning process and am very grateful that you opened this door for me that led to computational genomics area where I feel home now. Especially thanks to Sonja for putting time into making the paper and thesis as comprehensive and perfectly phrased as possible. I very much appreciate your steady effort and support!
This joyful and exciting survival has been possible thanks to the exchange and discussion with colleagues – especially Rohit, Bia, Amin, Shirley, Varo and Irma at Bioinf and many other people from the community at conferences and workshops. I thank Katja for "adopting" me as part of her group and letting me take part in the group meetings and journal clubs.
I would also like to thank my co-workers from the Informatics department in Leipzig, Nic and Tom, who carefully listened to my biological problems and magically converted them into algorithms. Thank you Alyssa for coming over to Leipzig twice, having this wonderful collaboration and becoming such a good friend. I love your spirit, girl! Thanks to Seva, the Gurevich and Meiler labs for making my stay in the US such a positive experience. Especially thank you Sergey for taking time and being so patient with me at my first steps back in the wet lab after three years!
This thesis would not have been possible without the administrative and technical support by Jens, Andrea and Petra, which goes far beyond the "normal" service.
Finally, there are the people outside of science that kept me sane in this intense time! Thanks to my family and friends, that keep me grounded. An dieser Stelle gibt es so viele Leute, denen ich danken möchte - den Biochemie-Mädels, den Restaurant-Testern und den Stadt-Mädels. Anni and Claudi, danke, dass ihr einfach da seid und immer ein offenes Ohr habt! Danke für die schöne Zeit, die wir immer zusammen verbringen - ob im Urlaub, bei einem Kaffee oder einer polnischen Rakete - danke dafür, dass ihr mich ab und zu aus meiner Wissenschaftswelt abholt und mit in die normale Welt nehmt.
Patrick - danke für deine Geduld und dafür, auch die schlecht gelaunte Version von mir zu ertragen und zu lieben! Ich kann es kaum glauben, doch nach diesem gefühlten Marathon haben wir jetzt tatsächlich beide den Uni-Lebensabschnitt gemeistert und alle Prüfungen und Arbeiten abgeschlossen!
Ein ganz besonderer Dank gilt meinen Eltern – für eure bedingungslose Unterstützung, Ermutigung und eure Ratschläge, die ich sehr schätze. Ohne eure Unterstützung stände ich heute nicht an dieser Stelle! Ihr seid die besten Eltern, die man sich vorstellen kann und habt mir die Neugier und das Durchhaltevermögen mitgegeben, die diese Arbeit erst ermöglicht haben.

Contents

Selbstständigkeitserklärung ii

Acknowledgements iv

1 Introduction 2
1.1 Basics and definitions . . . 2
1.1.1 What is a gene? . . . 2
1.1.2 What is a tree in phylogenetics? . . . 3
1.1.3 What are paralogs and orthologs? . . . 4
1.1.4 Central dogma in molecular biology: From DNA to protein . . . 5
1.2 Gene duplications as evolutionary playground . . . 12
1.2.1 Mechanisms of gene duplication . . . 13
1.2.2 Evolutionary fate of duplicated genes . . . 14
1.3 Identification and annotation of protein homologs . . . 15
1.3.1 Challenges of existing resources . . . 16
1.3.2 Similarity search approaches without consideration of the gene structure . . . 17
1.3.3 Gene structure aware gene annotation approaches . . . 19
1.3.4 Graph-based inference of orthology relationships . . . 21
1.3.5 Chance and challenge of fragmented assemblies . . . 21
1.4 Applied phylogenetic methods . . . 22
1.4.1 Phylogenetic inference in a nutshell . . . 23
1.4.2 Inference of natural selection in inter-species data sets . . . 29
1.4.3 Detection of specificity determining positions . . . 32
1.5 Multi-talents in cell signaling: The cytosolic arrestin proteins . . . 34
1.5.1 Functions of arrestins in cell signaling . . . 34
1.5.2 Arrestin activation by GPCR binding . . . 36
1.5.3 Functions of arrestins in cellular trafficking . . . 37
1.5.4 Evolution of arrestins . . . 39

2 The `ExonMatchSolver`-pipeline 42
2.1 Motivation . . . 42
2.2 Methods . . . 43
2.2.1 Pipeline overview . . . 43
2.2.2 Exon assembly as an assignment problem . . . 43
2.2.3 Solving the Paralog-to-Contig Assignment Problem . . . 46
2.2.4 Post-processing . . . 47
2.2.5 Implementation and usage . . . 48
2.2.6 Performance assessment by simulations . . . 50
2.3 Results . . . 50
2.3.1 Performance on simulated data . . . 50
2.3.2 Performance on real data - Two Showcase Examples . . . 51
2.4 Discussion . . . 57

3 Evolution of the arrestin protein family in deuterostomes 61
3.1 Motivation . . . 61
3.2 Material and Methods . . . 62
3.2.1 Database scan . . . 62
3.2.2 Detailed gene annotation . . . 63
3.2.3 Data resources used in the current study . . . 64
3.2.4 Alignment and building of phylogenetic trees . . . 64
3.2.5 Identification of specificity determining positions . . . 65
3.2.6 Testing for natural selection . . . 66
3.2.7 Assessement of conservation . . . 66
3.2.8 Parsimonious reconstruction of exon gain and loss events . . . 67
3.3 Results . . . 67
3.3.1 Evolution of the arrestin fold family based on database inquiries 67
3.3.2 The refined arrestin annotations are more complete than database entries . . . 72
3.3.3 Arrestin paralog gain and loss patterns based on the refined annotations . . . 73
3.3.4 Evolution of arrestin functional elements . . . 88
3.4 Discussion . . . 96
3.4.1 Limitation of arrestin database annotations . . . 96
3.4.2 Arrestins in early vertebrate evolution . . . 98
3.4.3 Sub- and neofunctionalization as consequence of the 3R-WGD . 102
3.4.4 Independent arrestin duplications in deuterostomes . . . 104
3.4.5 Loss of arrestin paralogs in different vertebrate orders . . . 106
3.4.6 Previously unknown interaction partners and isoforms . . . 108

4 Improvements on the `ExonMatchSolver`-pipeline 110
4.1 Motivation . . . 110
4.2 Methods . . . 111
4.2.1 Estimation of the paralog number . . . 111
4.2.2 Subdivision of gene loci on the same contig . . . 113
4.2.3 Implementation details . . . 113
4.2.4 Assessment of the `ExonMatchSolver`-pipeline Version `2` . . . 115
4.3 Results . . . 115
4.4 Discussion . . . 116

5 Conclusion and Outlook 119

A Additional figures 123

B Additional tables 134

C CV 152

Bibliography 156

List of Figures

1.1 Key concept of a gene. . . . 3
1.2 Hypothetical gene trees to clarify homology related terminology. . . . 4
1.3 Schematic depiction of the central dogma of molecular biology. . . . 5
1.4 Possible selective constraints acting on a gene. . . . 7
1.5 The genetic code. . . . 9
1.6 Schematic depiction of a profile hidden Markov Model. . . . 19
1.7 Missing data can lead to misassignment of paralogs and orthologs. . . . 29
1.8 Role of G proteins and arrestins in GPCR signaling. . . . 35
1.9 Functional elements of arrestins. . . . 38
1.10 Non-visual arrestins mediate endocytosis. . . . 40

2.1 Overview of the `ExonMatchSolver`-pipeline. . . . 44
2.2 Illustration of the paralog-to-contig assignment problem. . . . 45
2.3 Extended schematic of the `ExonMatchSolver`-pipeline. . . . 48
2.4 Accuracy and running time of the `ExonMatchSolver` on simulated data. . . . 52
2.5 Illustration of the paralog-to-contig assignment for arrestin paralogs in pufferfish. . . . 53
2.6 Phylogenetic tree of pufferfish and zebrafish arrestins. . . . 54
2.7 Illustration of the paralog-to-contig assignment for latrophilin paralogs in cod. . . . 56

3.1 Abundance of arrestin fold family members in animals. . . . 68
3.2 Abundance of arrestin fold family members in different domains of life. 70
3.3 Approximate maximum likelihood tree of arrestin fold family members. 71
3.4 Comparison of arrestin gene numbers in deuterostomes between the `ExonMatchSolver`-pipeline and `OrthoDB`. . . . 73
3.5 Unrooted maximum likelihood tree of arrestins. . . . 75
3.6 Rooted bayesian tree of arrestins. . . . 76
3.7 Duplication and deletion of arrestin paralogs in basal deuterostomes. . . . 78
3.8 Specificity determining positions discriminating between sea urchin *ARR0.1* and other *ARR0*s. . . . 80
3.9 Maximum likelihood tree of exon 5 sequences from arrestins of spotted gar, ghost shark and little skate. . . . 81
3.10 Duplication and deletion of arrestin paralogs within ray-finned fish. . . . 82
3.11 Temporal and spatial expression of arrestin genes in zebrafish. . . . 83
3.12 Specificity determining positions discriminating each pair of duplicated visual arrestins in teleosts. . . . 85
3.13 Duplication and deletion of arrestin paralogs in lobe-finned fish. . . . 86
3.14 Synteny of the *ARR3* locus in afrotherians and xenarthrans. . . . 87
3.15 Changes in conservation patterns and functional motifs of arrestins. . . . 89
3.16 Evolutionary changes in exon–intron structure of arrestins. . . . 91

3.17 Alignment of exon–intron borders after insertion of intron 85c into exon 5. . . . 92
3.18 Conservation pattern of the minor clathrin binding site in arrestins. . . . 94
3.19 Conservation pattern of the Adapter Protein-2 motif in arrestins. . . . 95
3.20 Putative effect of arrestin truncation by use of an alternative start codon within exon 8. . . . 96
3.21 Putative effects of conserved splice variants on structure and binding interfaces of arrestins. . . . 97

4.1 Overview of the ExonMatchSolver-pipeline Version 2. . . . 114
4.2 Comparison of different Versions of the ExonMatchSolver-pipeline predicting arrestin genes in the purple sea urchin genome. . . . 116
4.3 Parameters of ExonMatchSolver solutions in dependence on the paralog number. . . . 117

A.1 Scan of the UniProtKB with arrestin profile Hidden Markov Models employing jackhmmer. . . . 123
A.2 Foreground branches in the natural selection analysis of arrestins. . . . 124
A.3 Approximate maximum likelihood tree of the single arrestin domain hits. . . . 125
A.4 Abundance of arrestin fold family members in bilaterians. . . . 126
A.5 Rooted bayesian tree of arrestins. . . . 127
A.6 Unrooted maximum likelihood tree of partial arrestins. . . . 128
A.7 Arrestin paralogs within laurasiatherians and superprimates. . . . 129
A.8 Summary of arrestin gene, exon and intron gain and loss events in deuterostomes. . . . 130
A.9 Candidate loci and genes for *ARR3* in armadillo. . . . 131
A.10 Genomic locus of the *ARR3* pseudogene in shrew. . . . 131
A.11 Expression pattern of *SAG* and *ARRB1* in opossum. . . . 132
A.12 Structure and genomic locus of the *ARRB1.2* retrogene in wallaby. . . . 132
A.13 Pfam domains in deuterostome arrestins. . . . 133

List of Tables

1.1 Reversible post-translational modifications. 12
1.2 Parameters in branch-site models of natural selection used in the current study. 31
1.3 Key functional elements in arrestin activation and receptor binding. . . 37

2.1 Performance of `Scipio` and the `ExonMatchSolver`-pipeline in prediction of arrestin genes in pufferfish. 55
2.2 Performance of `Scipio` and the `ExonMatchSolver`-pipeline in prediction of latrophilin genes in cod. 58

3.1 Scan of databases with arrestin profile Hidden Markov Models. 69
3.2 Position of lamprey arrestins in phylogenetic inferences. 77
3.3 Positively selected residues of arrestins. 79

4.1 Type assignments of the `ExonMatchSolver` Version `2` implementation. 118

B.1 List of genomes considered for a refined annotation of arrestins. 134
B.2 Cross-paralog conservation of arrestin isoforms according to public resources. 137
B.3 List of additional omics data considered for a refined annotation of arrestins. 139
B.5 Selection pressure acting on positively selected foreground branches of arrestin gene trees. 139
B.7 Specificity determining positions identified for different arrestin subgroups. 140
B.8 Functional residues of arrestins considered in the current study. 141
B.4 Model selection during Bayesian inference of arrestin trees. 143
B.9 Fragments of *ARRB2* detected in birds. 143
B.6 Analysis of natural selection after arrestin duplication. 146
B.10 Arrestin residues with post-translational modifications. 147
B.11 Pattern and conservation of arrestin post-translational modifications. . 149
B.12 Cross-paralog conservation of arrestin isoforms. 151

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
AP-2	**A**dapter **P**rotein-2 Complex
BEB	**B**ayes **E**mpirical **B**ayes Method
BF	**B**ayes **F**actor
BG	**B**ack**g**round Branch
BIC	**B**ayes **I**nformation **C**riterion
BLOSUM	**Blo**cks **Su**bstitution **M**atrix
BS	**B**ootstrap **S**upport
CBS	**C**lathrin **B**inding **S**ite
ChIP-Seq	**Ch**romatin **I**mmuno**p**recipitation DNA-**Seq**uencing
cDNA	**c**omplementary **D**eoxyribo**n**ucleic **A**cid
DNA	**D**eoxyribo**n**ucleic **A**cid
EMS	**E**xon**M**atch**S**olver
ENCODE	**Enc**yclopedia **O**f **D**NA **E**lements
EST	**E**xpressed **S**equence **T**ag
FG	**F**ore**g**round Branch
GDP	**G**uanosine **Di**phosphate
GPCR	**G** Protein-**C**oupled **R**eceptor
GRK	**G** Protein **R**eceptor **K**inase
GTR	**G**eneral **T**ime **R**eversible substitution model
GTP	**G**uanosine **Tri**phosphate
HKY	**H**asegawa, **K**ishino and **Y**ano nucleotide substitution model
HMM	**H**idden **M**arkov **M**odel
HSP	**H**igh-scoring **S**egment **P**air
ILP	**I**nteger **L**inear **P**rogramming
IP6	**I**nositol-**hexa**-**p**hosphate
JC69	**J**ukes **C**antor 1969 substitution model
JTT	**J**ukes **T**aylor **T**hornton substitution model
LINE	**L**ong **In**terspersed **E**lements
LG	**L**e **G**ascuel substitution model
LRT	**L**ikelihood **R**atio **T**est
MC	**M**arginal **C**onserved
MCA	**M**ultiple **C**orrespondence **A**nalysis
MCMC	**M**arkov **C**hain **M**onte **C**arlo
ML	**M**aximum **L**ikelihood
MSA	**M**ultiple **S**equence **A**lignement
mRNA	**m**essenger **R**ibo**n**ucleic **A**cid
nc	**n**on-**c**oding
NNI	**N**earest **N**eighbor **I**nterchange
NP hardness	**N**ondeterministic **P**olynomial time complexity of a decision problem
PAM	**P**oint **A**ccepted **M**utation matrix
PCAP	**P**aralog-to-**C**ontig **A**ssignment **P**roblem
PDB	**P**rotein **Dat**a**b**ase

pHMM	**p**rofile **H**idden **M**arkov **M**odel
pre-mRNA	**Pre**cursor **m**essenger **R**ibo**n**ucleic **A**cid
PTM	**P**ost-**T**ranslational **M**odification
RNA	**R**ibo**n**ucleic **A**cid
RNA-Seq	**R**ibo**n**ucleic **A**cid-**Seq**uencing
SH	**S**equence **H**armony
SDP	**S**pecificity **D**etermining **P**osition
SPR	**S**ubtree **P**runing and **R**egrafting
SRA	**S**hort **R**ead **A**rchive
SS	**S**plice **S**ite
SUMO	**S**mall **U**biquitin-related **Mo**difier
TCE	**T**ranslated **C**oding **E**xon
TF	**T**ranscription **F**actor
tRNA	**t**ransfer **R**ibo**n**ucleic **A**cid
TSS	**T**ranscription **S**tart **S**ite
pp	**p**osterior **p**robability
UTR	**Un**translated **R**egion
Vps26	**V**acuolar **p**rotein **s**orting-associated protein 26
WAG	**W**helan **a**nd **G**oldman substitution model
WGD	**W**hole **G**enome **D**uplication
WT	**W**ild **T**ype

aa	**a**mino **a**cid
B	**B**ytes
b	**b**ases
bp	**b**ase **p**air
Da	**Da**lton
my	**M**illion **y**ears
mya	**M**illion **y**ears **a**go
nt	**n**ucleo**t**ide

Gene names are used in accordance to the HUGO gene naming convention and are not given in the abbreviation list. Furthermore, the one letter code is used to denote amino acids and the letters A, C, G, T, U, Y (C or T) to denote nucleotides. Mainly in the Appendix, three to four letter codes are used to abbreviate species names (please see Tab. B.1 for the respective abbreviations).

Chapter 1

Introduction

1.1 Basics and definitions

This first section is directed to the reader who is not familiar with basic concepts of molecular (computational) biology, such as phylogenetic tree, gene, information flow from DNA to protein and homology. Please refer to an introductory biochemistry book if you are not familiar with the concepts of deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. The central dogma will be explained with a strong focus on the different constraints and pressures that act on structure and sequence of a protein-coding gene's DNA, RNA and protein molecules.

1.1.1 What is a gene?

The definition of gene has undergone changes during recent years with the definition discussed below been inspired by the proceedings of the human Encyclopedia of DNA elements (ENCODE) consortium, which aims to provide a comprehensive annotation of functional elements. Until the early 2000s, the accepted definition of a gene has been a heritable unit that connects the phenotype with the genotype. Different phenomena such as imprinting, epigenetics, RNA-editing, protein modifications and protein splicing are now known to influence sequence or structure of the functional product and thus complicate this definition. In a revision of the gene definition, Gerstein et al. (2007) defined a gene as a union of genomic sequences that encode a coherent set of potentially overlapping functional products. The genomic sequence is encoded by DNA with few exceptions of genomes consisting of RNA (e. g. RNA viruses). From the structural point of view, a gene is composed of exons and introns. During gene expression, a DNA-encoded gene is transcribed into RNA, which is subsequently processed and can give rise to many different versions (isoforms) of the same gene. Exons are those genomic sequences, that are included in the processed transcript, while introns are transcribed, but usually excluded in a process called splicing. The hypothesis about the ensembl of differently structured gene transcripts arising from a single gene is called gene model. In the context of this thesis, the gene model is called "gene structure" or "exon–intron structure" as the thesis focuses on a single isoform per gene unless specified otherwise.
Parts of the processed transcript can subsequently be read in units of three nucleotides (a codon) and translated into an amino acid sequence. The region of the protein-coding gene that is not translated, but part of the processed/spliced transcript is called untranslated region (UTR). Those genes are called protein-coding genes, while genes that encode transcripts that are not translated are called non-coding (nc) genes. The two different possible classes of functional products, protein and ncRNA, are considered separately in the gene definition. Let's illustrate the definition of a gene with the help of some examples (Fig. 1.1). Gene #1 is considered as a single gene

although encoding three different spliced transcripts as (1) the exons of one spliced transcript overlap with at least one other spliced transcript encoded by the same gene locus and (2) all transcripts encode the same class of functional product (here: protein). The strict consideration of the functional level results in the exclusion of regulatory elements from the gene definition, as they can regulate the expression of more than one gene (gene #2 vs. gene #3 in Fig. 1.1).

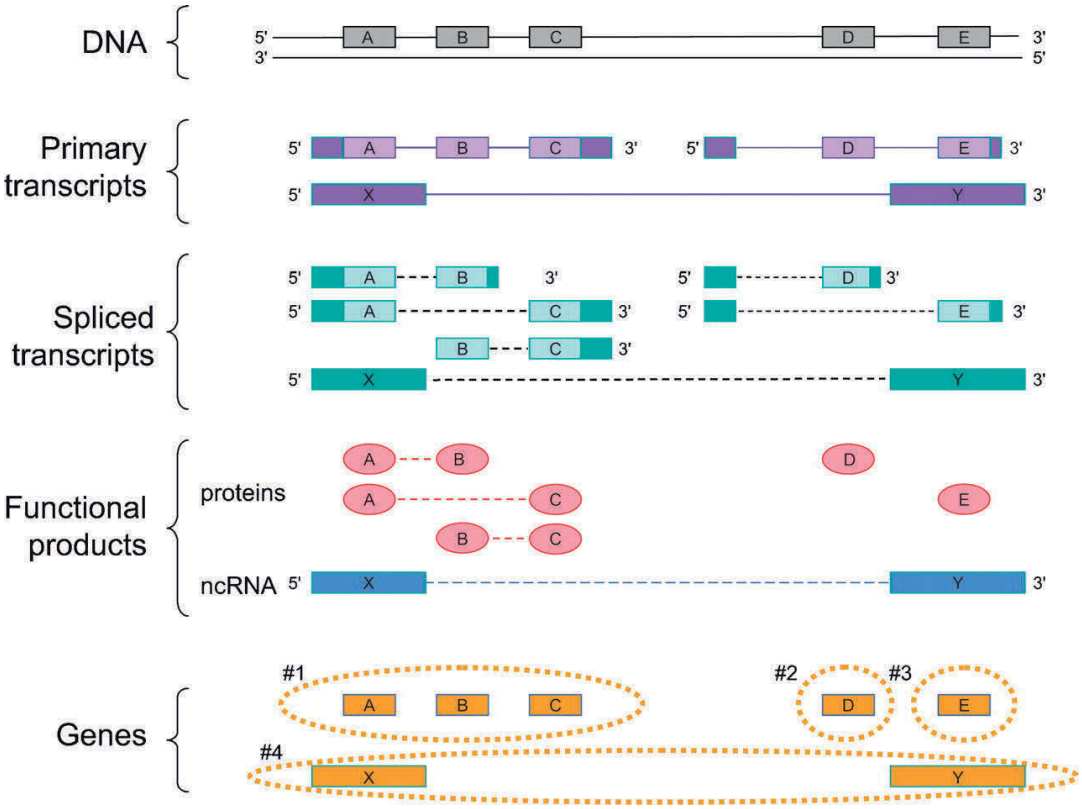

Figure 1.1: Key concept of a gene. The genomic locus encodes three primary transcripts (purple). Two of these encode proteins (first line), one a ncRNA (second line). Only the protein-coding parts of the respective exons are projected onto the DNA. Processing of the primary transcripts increases the number of functional products to six. Three of those (left side) share at least one exon with another transcript of the same functional class (here protein) resulting in a total gene count of four. Although gene #4 shares genomic sequence with genes #1 and #3, these are considered separately as they encode molecules of different functional classes. Untranslated regions (UTR) are shown in dark color. The figure was taken from Gerstein et al. (2007).

1.1.2 What is a tree in phylogenetics?

A tree is an acyclic, connected graph with g vertices (or nodes, leaves) and h edges (or branches). In this work I am mostly concerned about directed, rooted trees (e. g. in Fig. 1.2). The root node is a special, labeled vertex in a time-directed tree with the degree two. Interior nodes have a degree of three or more, while leaves possess the degree one. If a tree contains internal nodes with a degree greater than three, this node is a multi-furcation and the tree is not fully resolved. Unrooted trees can be fully described by a set of bipartitions (splits) along the tree's edges.
In phylogenetics, trees are mainly built from three kinds of data, the amino acid alphabet (protein: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y), the nucleotide

alphabet (DNA: A, C, G, T, RNA: A, C, G, U) and the presence-absence pattern of genomic features, such as genes or introns. Taxonomists are usually interested in resolving a species' position in the tree of life in relation to other known species thus answering questions like which rank, family or order the species belongs to. They are thus interested in species trees, where every node represents a speciation event.
I am concerned with a different kind of tree, the gene tree, where nodes represent either duplication or speciation events. Gene trees are generated from amino acid or nucleotide sequences of genes or gene parts, which may be identical to the species tree. A gene tree is described by its tree topology and its branch lengths, which approximate the amount of evolutionary changes along the branch, called divergence. The branch lengths of a gene tree are usually expressed in units of differences or substitutions along the branches. The substitution rate normalizes the total number of substitutions over time.

1.1.3 What are paralogs and orthologs?

Homology describes the relationship between two characters (e. g. genes) that descend from a common ancestor. If two characters share similarity that arose by convergence and not by descendance, this relationship is referred to as analogy (Fitch, 2000). Analogy is a frequent characteristic in structural biology, where the substructure of a protein is conserved although the respective molecules are not homologs (Illergård, Ardell, and Elofsson, 2009). On molecular level, two kinds of homology exist: paralogy and orthology. Orthologs descend from a speciation event and subsequent divergence (Fig. 1.2). By definition, the character has an ortholog in a different species. Paralogs arise from a duplication event. If a speciation event follows the duplication, these characters are called out-paralogs (Fig. 1.2 A). In-paralogs are always in the same species as no speciation followed the duplication event (Fig. 1.2 B, Fitch (2000)).

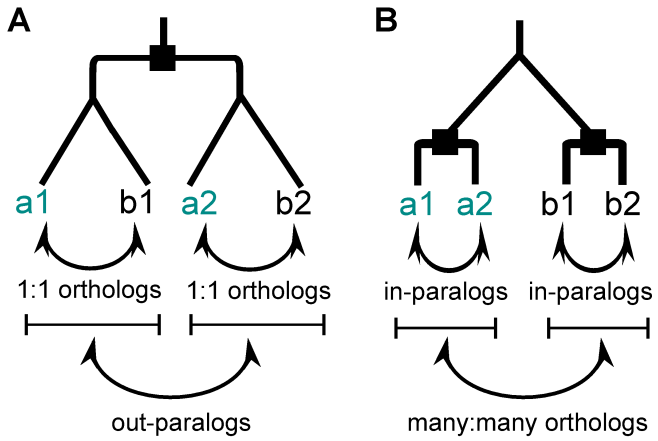

Figure 1.2: Hypothetical gene trees to clarify homology related terminology. Two possible gene tree topologies are shown for a set of two genes in each of two species (A, B). The letters a and b within the tree represent different species, e. g. armadillo (light blue) and brown bear (black), while 1 and 2 are names for genes that arose by gene duplication. Gene duplications are indicated by a square and a horizontal bifurcation, while speciations are indicated by sloping lines.

1:1 orthologs have a 1:1 relationship in the different species of interest. Duplication events after the speciation lead to 1:many or many:many orthology relationships (Fig. 1.2 B). Paralogs frequently have different functions (are non-isofunctional), while 1:1 orthologs are more often isofunctional (Ohno, 1970; Altenhoff et al., 2012).

1.1.4 Central dogma in molecular biology: From DNA to protein

Protein expression is a complex multi-step process within the cell that can be controlled on different levels to result in temporal and cell-specific expression regulation. The central dogma in molecular biology describes the information flow from the genomically encoded information of a gene (DNA) to an information copy that is transported out of the nucleus (RNA) to be translated into amino acid sequences, which can fold into a functional protein (Fig. 1.3). Each of these processes poses sequence and structural constraints on one of the three primary levels: protein → RNA → DNA. Because of the nature of information flow, primary constraints on one level will be reflected as constraints on the lower levels, e. g. constraints on protein structure will be reflected on RNA and DNA level. The biological main implications of my thesis are based on making conclusions about protein sequence and function based on genomic information encoded in DNA. As the focus of this work is on protein-coding genes, I will assume a gene to be protein-coding in the following and exclude nc genes from consideration.

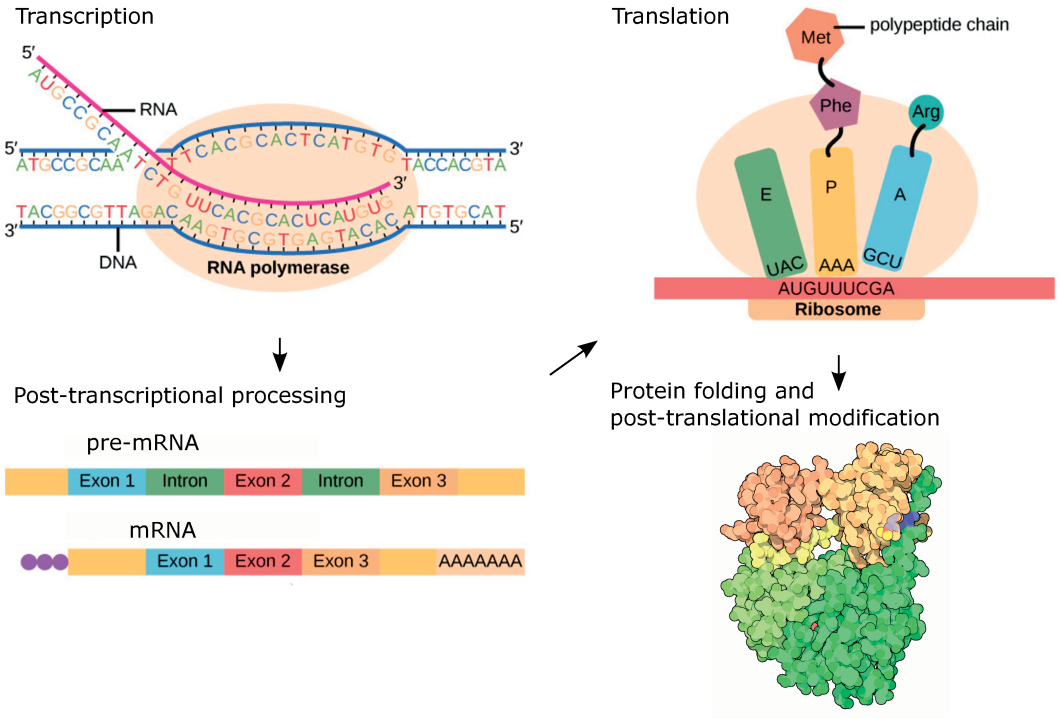

Figure 1.3: Schematic depiction of the central dogma of molecular biology. During protein expression, information encoded on DNA is transcribed into a RNA molecule, which is translated into a protein. The pre-mRNA molecule is processed including the addition of a 5′ cap (violet) and a poly-A tail and the removal of introns (splicing). The exemplary protein shown here, protein tyrosine kinase c-Src, is modified after translation (phosphorylation marked in blue). Only the protein-coding part of the exons are shown in bright colors, while the untranslated regions of exons 1 and 3, respectively, are shown in yellow. The figure is based on OpenStax CNX (2017-09-13) and Goodsell (2003).

Transcription

The human genome encodes about 21,000 protein-coding genes (The ENCODE Consortium, 2012). The result of transcription is a copy of the protein-coding gene written

in RNA (called precursor messenger RNA, pre-mRNA, Fig. 1.3). Transcription is initiated by binding of the RNA polymerase II to the DNA in a sequence-dependent manner. The sequence stretch, where the polymerase binds, is called promotor and situated towards the genomic 5′ end relative to the initiation point of transcription, the transcription start site (TSS, Alberts (2011)). The TSS is located at position +1 by convention and will be used as a reference point to refer to positions situated upstream (towards the 5′ end) or downstream (towards the 3′ end).
Transcription initiation, elongation and termination is strongly influenced by the presence and interaction of the RNA polymerase II with proteins (transcription factors, TF) and RNA-molecules (e. g. long ncRNAs). TFs recognize and bind specific DNA sequence motifs of about 6-20 nt length. Sequence stretches (binding sites) that promote the function e. g. transcription are called enhancers, while repressing sequences are called silencer. Although TFs primarily bind within several hundred nucleotides upstream of the TSS (Koudritsky and Domany, 2008), TF binding sites can also be located several 100 kb apart from the TSS (Fig. 1.4). Stergachis et al. (2013) demonstrated recently, that at least 14 % of all protein-coding bases in human have contact to a TF in at least one cell type with the majority of bases located in the first exon. TF recognition thus poses constrains on the codon and eventually amino acid choice of these “dual-use codons”.
Due to the short length of the binding motifs, every TF can bind up to several 1,000 locations within the genome just by chance. The occupancy of individual binding sites is the outcome of a complex interplay between concentration and binding affinity of different TFs, which is often cooperative. Moreover, DNA packing and accessibility contribute to cell-type specific and temporal regulation of transcription. This regulation can result in the usage of alternative TSSs. Resulting transcripts might differ in their 5′ UTR or 5′ protein-coding sequence (1, 2 vs. 3, 4 in Fig. 1.4) thus contributing to the diversity of different functional molecules on RNA and protein level.

Post-transcriptional processing and splicing

As part of a quality control system, the pre-mRNA transcript is processed before translation. This process encompasses 5′ capping, 3′ cleavage and polyadenylation as well as splicing (Fig. 1.3, 1.4). Post-transcriptional processing happens in parallel to transcription and has been observed to influence transcription and vice versa (Jonkers and Lis, 2015). Furthermore, RNA-editing can change the primary sequence of the pre-mRNA potentially implicating any of the following layers of gene regulation, e. g. splicing or the identity of the encoded amino acid (Fritzell et al. (2017) and references therein).
As RNA has a much more flexible backbone than DNA, function mediated by structure gains importance. RNA structural elements like riboswitches and RNA thermometers regulate transcription and translation by blocking or freeing the TSS or translation start site depending on ligand concentration and temperature, respectively (Wachter (2014) and references therein).
Splicing leads to the excision of intron sequences from the pre-mRNA molecules. The spliceosome, the RNA-protein complex, that catalyzes the splicing reaction recognizes sequence motifs called 5′ splice site (SS) and 3′ SS located at the very end of the respective intron as well as an A within the intron sequence, the branch site. The spliceosome catalyzes a two step reaction: (1) Bond formation between the A branch point and the 5′ SS forming a lariat structure, where the 5′ SS is no longer

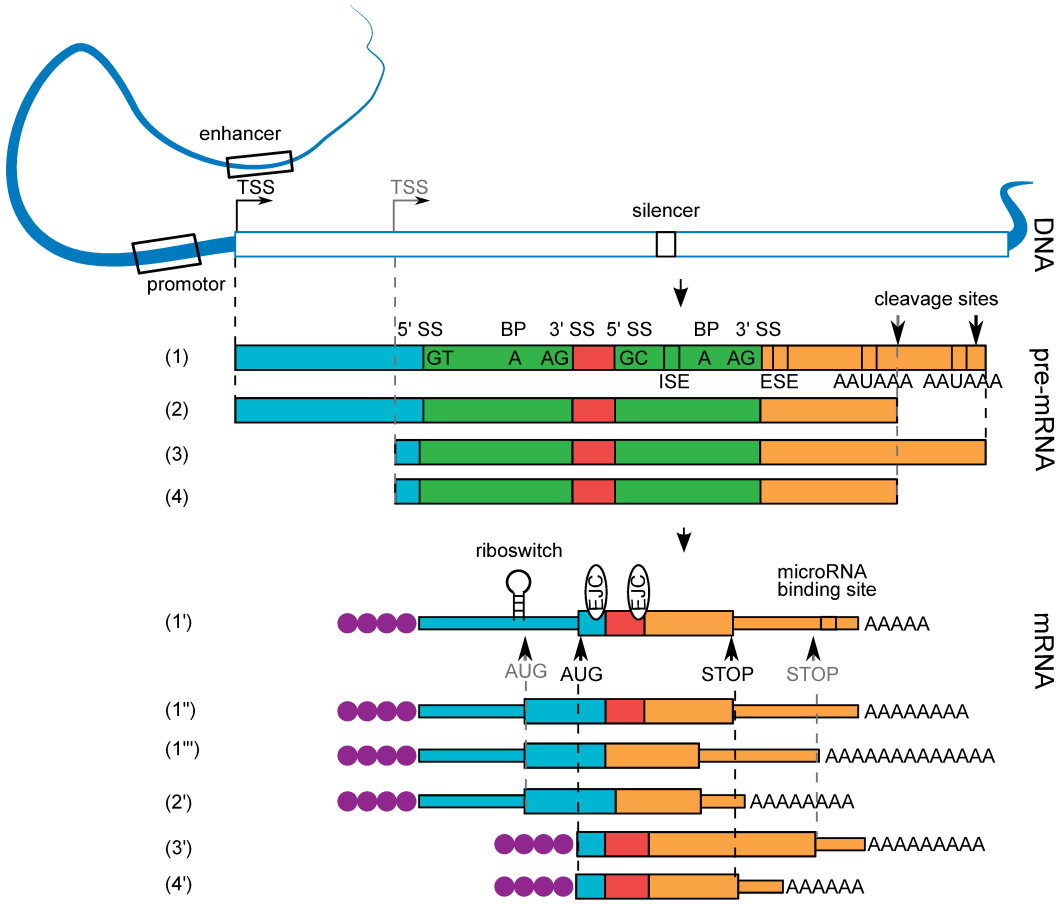

Figure 1.4: Possible selective constraints acting on a gene. Selective constraints on a gene arise from every information layer, DNA, RNA and the protein sequence. For simplicity, only a selection of constraints arising from RNA and DNA sequence (e. g. enhancer and silencers) and structure (e. g. riboswitch, DNA structure) are shown. Parts of the mRNA that will not be translated in a polypeptide chain (untranslated regions or UTR) are shown as thinner lines. Please note that sequence motifs are depicted by the one letter nucleotide code or simplified by boxes. Alternative transcription and translation start and stop sites are shown in grey. Abbreviations: BP – branch point; ESE – exonic splicing enhancer; ISE – intronic splicing enhancer; EJC – exon junction complex; SS – splice site; TSS – transcription start site.

connected to the pre-mRNA, i. e. it is cleaved; (2) Cleavage of the 3′ SS from the pre-mRNA and exon ligation (Alberts, 2011).

The composition of the spliceosome can differ in regard to its RNA components giving rise to two canonical types, the U2 and U12 spliceosome. Both types slightly differ in their sequence preferences for the 5′ and 3′ SSs. The majority of SSs in human, 98.9 %, have a GT-AG signature, followed by GC-AG (0.88 %, Parada et al. (2014)). The SSs as denoted here, are part of the intron. The sequence preference in fact extends over these two nucleotides into the intronic and exonic sequences, although being less important. Most U2/U12-like non-canonical SSs (not GT-AG) represent an alternative to a canonical SS (Parada et al., 2014). The selection of some SSs over others might result in intron retention, exon skipping or exon inclusion (1″ vs. 1‴ in Fig. 1.4, Sibley, Ule, and Blazquez (2016) and references therein).

Splicing is highly regulated by splicing factors, which bind to exonic or intronic splicing enhancers and silencers in a tissue-specific manner (Wang et al., 2006; Wainberg,

Alipanahi, and Frey, 2016). In case of alternative exons, splicing factors preferentially bind up to 300 nt upstream or downstream of the respective exon or within the exon itself (Barash et al., 2010). Other mostly sequence dependent post-processing steps are 3′ cleavage and polyadenylation. Tissue-specific alternative cleavage and polyadenylation can influence the 3′ coding sequence and the 3′ UTR (Fig. 1.4, 1′, 1″, 1‴ vs. 2′). The cleavage and polyadenylation specificity factor complex binds to a 6 nt key motif (AAUAAA) that is situated about 15-30 nt upstream of the cleavage site (Elkon, Ugalde, and Agami, 2013). Binding of different cleavage factors to U-rich downstream and upstream sites can further enhance cleavage (Fig. 1.4).
The processed mRNA molecule is finally transported to the cytoplasm for translation at the ribosomes. The transport depends on nuclear export factors of the TREX complex, which are recruited to the mRNA molecules through interactions with proteins of the transcription and splicing machinery close to the 5′ end of the mRNA (Masuda et al., 2005; Cheng et al., 2006). Some components of the TREX complex (CHTOP, ALY, NXF1) directly interact with the mRNA (Viphakone et al., 2012; Chang et al., 2013).

Translation

During translation, the coding part of the mRNA is re-written into an amino acid sequence (Fig. 1.3). The common basic unit of translation is a triplet of nucleotides, a codon. Each of the 64 possible codons (4^3) encodes one of 21 amino acids or a stop signal (UAA, UAG or UGA) in vertebrates (Vertebrata). Due to the degeneracy of the code, several codons can encode the same amino acid. Usually, the third position is most variable (Fig. 1.5).
Transfer RNAs (tRNAs) are the transport vehicles of amino acids, coupling amino acid and codon by displaying a triplet complementary to the codon (anti codon). The ribosome, a huge RNA-protein complex, catalyzes translation in three steps (upper right corner in Fig. 1.3): (1) Binding of the loaded aminoacylated-tRNA to the codon (position A); (2) Transfer of the nascent peptide sequence to the aminoacylated-tRNA by formation of a peptide bond (position P); (3) Release of the empty/unloaded tRNA molecule by a downstream sliding movement of the ribosome (position E). The result of translation is a covalently linked sequence of amino acids that is dictated by the RNA sequence. Translation is initialized by the initiator-tRNA binding to the start codon (AUG), which is surrounded by a consensus sequence. The first amino acid of the nascent peptide is thus always a methionine. Skipping of a translation start sites can lead to translation initiation at alternative, downstream sites (Fig. 1.4, 1′ vs. 1″). If the reading frame is kept, this leads to an alternative protein N-terminus, while an alternative reading frame can result in a completely different protein sequence although exons are shared. Translation terminates when a stop codon is reached. The sliding of the translation machinery along a newly transcribed mRNA molecule results in the removal of all passed exon junction complexes, which mark former SSs. An mRNA molecule can be translated several thousand times depending on its half-life. Unsurprisingly, codon usage is constrained by optimization of transcription speed, translation elongation rates and influences co-translational protein folding (Zhou et al., 2016; Quax et al., 2015).
Several translation regulators (proteins or ncRNAs) bind to the mRNA molecule in a sequence and/or structure-dependent way. Especially, the role of *trans*-acting ncRNAs (long ncRNAs, microRNA, small interfering RNA) in (mis)regulation of translation, mRNA stability and decay has come to age during the last decades (Derrien, Guigó, and Johnson, 2011; Cloonan, 2015). Translation is also regulated

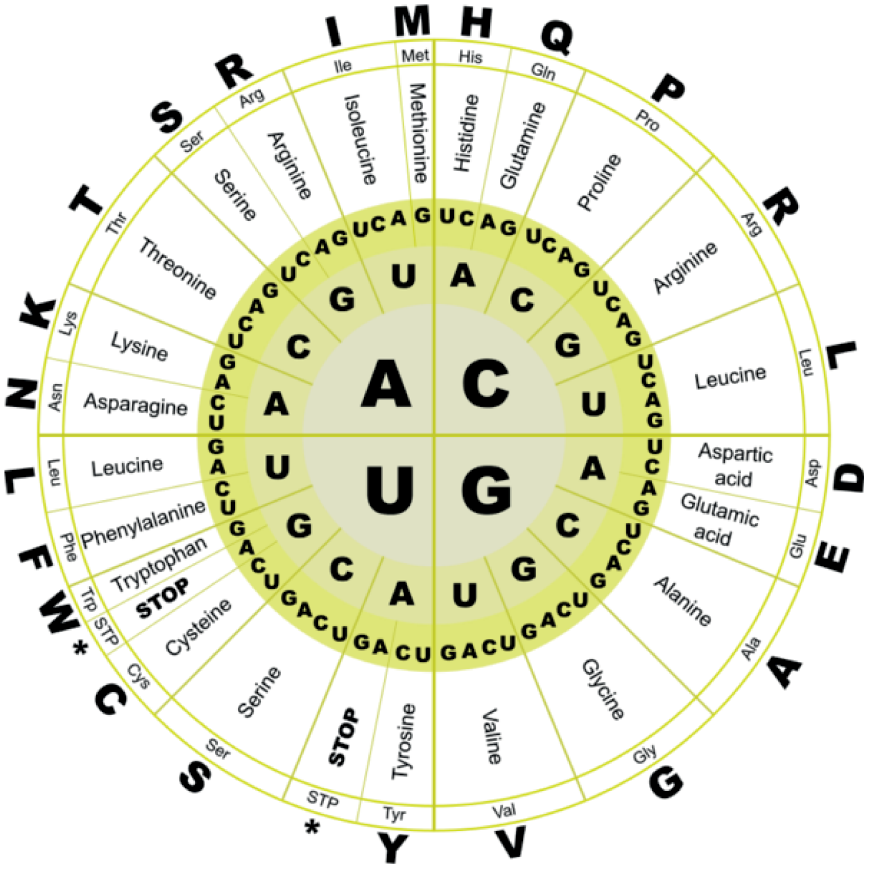

Figure 1.5: The genetic code. Three RNA nucleotides (a triplet or codon) encode one amino acid with the first codon position shown in the most inner yellow wheel, the second in the middle wheel and the third codon position in the outer wheel, respectively. The three letter amino acid code used throughout the manuscript and the one letter code, which is the common representation of amino acids as single characters for computational methods are depicted. The genetic code is degenerative, meaning that one amino acid is encoded by more than one codon. The last codon position is often variable without changing the identity of the encoded amino acid. Special codons encode a stop signal. Figure taken from http://images.clipshrine.com/download/wheel/large-Genetic-Code-RNA-0-44276.png, 2017/11/13.

by *cis*-acting RNA elements such as riboswitches, RNA thermometers or internal ribosome entry sites, that are mainly located in the 5′ UTR. mRNAs that bind an exon junction complex 50 nt downstream of a stop codon are usually degraded in a process denoted as non-sense mediated decay. This quality control mechanism usually ensures degradation of mRNAs with a premature stop codon that might result from intron retention or exon inclusion. Furthermore it opens up another layer of regulation as intron retention and exon exclusion may cause mRNA degradation unless inhibited (Lykke-Andersen and Jensen, 2015).

Protein folding

The lowest energy conformation or fold of a protein is mainly determined by biophysical constraints imposed by the amino acid sequence. Non-covalent interactions (van der Waals interactions, hydrogen bond bridges, electrostatic interactions) and covalent bonds other than the peptide bond (backbone), e. g. disulfide bonds between

Cys residues add to the energy gain in comparison to the unfolded state. Two of those non-covalent interactions – van der Waals interactions and hydrogen bonds – result in the burying of hydrophobic residues into the protein core, while polar amino acids satisfy their hydrogen bonding potential with water molecules from the aqueous environment (the solvent). The formation of the hydrophobic core is often the main driver of protein folding and one of the main predictors of the amino acid substitution rate of the specific sites (Echave, Spielman, and Wilke (2016), referred to as site-specific rate in the following). Solvent-inaccessible sites evolve much slower than residues on the protein surface (Goldman, Thorne, and Jones, 1998; Lin et al., 2007; Franzosa and Xia, 2009). Another related, powerful predictor of site-specific rates is the contact packaging or contact density as measured by the weighted contact number (Echave, Spielman, and Wilke (2016), all residues weighted by their inverse square root of their distance to the residue of interest). More densely packed residues evolve slower than residues with fewer partners in their neighborhood (Yeh et al., 2014a; Yeh et al., 2014b).

Formation of hydrogen bonds between peptide backbones dictated by the backbone's possible rotation angles (Ramachandran, Ramakrishnan, and Sasisekharan, 1963) give rise to regular secondary structure elements, namely the α-helix (Pauling, Corey, and Branson, 1951) and β-sheets (Pauling and Corey, 1951). As hydrogen bonds are formed between the peptide backbone, α-helices and β-sheets form independently of the identity of the amino acid side chains and are re-occurring structure elements in most proteins. Although the secondary structure has low predictive power of site-specific rates in comparison to solvent-accessibility (Goldman, Thorne, and Jones, 1998), hydrogen bond formation between the peptide backbone and between the peptide backbone and amino acid side chains are a key constraint in protein folding and have an effect on amino acid conservation (Worth, Gong, and Blundell, 2009). Apart from those structured regions, proteins can contain or be completely composed of disordered regions, that generally evolve faster (Brown, Johnson, and Daughdrill, 2010; Brown et al., 2002), although selection on intrinsically disordered proteins is not well understood (Chi and Liberles, 2016). α-helices and β-sheets are itself part of a limited set of hydrogen bonding favorable super-secondary structures or folding motifs, such as coiled-coiled helices, β-sandwich, β-barrels, β-propellers and jellyrolls (Worth, Gong, and Blundell, 2009). More complex, but conserved, stable protein substructures that consist of a fixed arrangement of secondary structure elements and often fold independently are called protein domains. Many proteins are composed of several domains that are connected by flexible loop regions. While many protein domains are conserved across all domains of life, the conservation of specific combinations seems to be more restricted (Lees et al. (2016) and references therein). As a consequence of the functional and structural importance of domains, domains/folds are usually more conserved than amino acid sequences. Although the individual amino acid sequence dictates a protein's lowest energy conformation, this conformation is usually not the most stable thermodynamic structure accessible through individual amino acid substitutions from an ancestral sequence. Instead, most proteins are marginally stable with a free energy just above the unfolded state (Pace and Hermans, 2008). In contrast, alternative conformations with a similar stability, that might be kinetically accessible during the folding process, have been observed and modeled to be selected against (Noivirt-Brik, Horovitz, and Unger, 2009; Minning, Porto, and Bastolla, 2013).

Within the cell, the amino acid chain is successively released from the ribosome. Local secondary structure elements that might connect to domains are formed first. While this is a rather fast process, local conformational changes of side chains and backbone

conformations frequently require more time. Even in the native state, the folded protein will naturally change between different conformations. Although certain protein conformations are necessary to perform specific functions, it is not known whether those states evolve neutrally or are selected for (Chi and Liberles, 2016). In general, evolution seems to favor fast folding and might select against kinetic traps that could e. g. arise by the presence of rare codons (Kimchi-Sarfaty et al., 2007).
A protein usually does not exist in isolation, but is in contact with other molecules (metabolites, ligands, proteins, DNA, RNA) that co-exist in the cell. Some of these interactions are key for the protein's function and will be selected for in regard to specificity and affinity, e. g. interaction with specific ligands or substrates. Amino acid chains from the same or from different proteins can interact forming homo- or hetero-dimers or higher molecular structures (oligomers). These interactions usually lead to a decrease in the site-specific rates of contacting residues as well as of some residues that are not situated in the immediate neighborhood (Dean et al., 2002). Substitutions on interaction interfaces of one protein might be compensated by substitution of contacting residues of the interaction partner, termed co-evolution. Furthermore, substitutions will be avoided that result in unwanted protein-protein interactions or binding of unwanted ligands (Chi and Liberles, 2016). In general, the expression level of a protein will highly influence the selection process with stronger selection acting on highly expressed proteins. Among other reasons, this is caused by a higher population of all conformations increasing non-specific interactions (Levy, De, and Teichmann, 2012). Furthermore, translation mistakes will occur more frequently resulting in possibly misfolded or aggregated proteins. Wilke and Drummond (2006) suggested a selection "for translational robustness" acting on highly expressed genes making respective proteins more robust in regard to the effect of missense substitutions on the overall fold.
Taking the different constraints and layers of selection on proteins into account, it is not surprising that only few protein properties are optimal in the space of all possible amino acid sequences. A specific position within the sequence space accessed during evolution rather depends on the ancestral sequence as substitutions strongly depend on the protein context (Alexander et al., 2009; Lunzer et al., 2010).

Post-translational modifications

Post-translational modifications (PTM) represent an additional regulatory layer that enables modification of protein properties after translation is completed. PTMs enable the cell to adapt towards changes in the environment in a much shorter time than needed for protein expression. All temporary modifications change the biochemical property of the respective amino acid side chain or backbone and thus have the potential to change protein stability, folding and activity (Bürkle (2002), Tab. 1.1). The modifying enzymes recognize the to-be-modified amino acid and neighboring amino acids in their substrate in a sequence-dependent manner, that might require the previous action of another modifying enzyme adding a combinatorial component to the process (Alberts, 2011). The most common PTM, phosphorylation, results in a change of charge and space requirement. Phosphorylation makes the modified residues, threonine, serine and tyrosine in eukaryotes, more hydrophilic. The modification is added by an enzyme class called kinase, while a phosphorylation is removed by a phosphatase. Phosphorylations are especially common in signal transduction cascades, where phosphorylation of a protein leads to the recruitment of a kinase that phosphorylates a downstream substrate. Those scaffolding proteins

have a crucial role in providing the spatial connection between the modifying enzyme, upstream regulators and downstream substrates. The activity of kinases itself is frequently regulated by phosphorylation, i. e. the kinase's substrates are other kinases (Fig. 1.3, e. g. in the MAP kinase pathway, Johnson and Lapadat (2002)). Other important, reversible PTMs are ubiquitination and SUMOylation. Ubiquitin and the small ubiquitin-related modifier (SUMO) are similar small peptides of 76 and 100 aa length, respectively, that are attached to a lysine residue in a multi-step process. Attachment of ubiquitin or SUMO has the potential to profoundly change the proteins structure and interaction with other proteins. Mono-ubiquitination regulates the internalization of trans-membrane receptors into the cell (endocytosis) and sorting, while poly-ubiquitination is a degradation signal attached to partially unfolded proteins (Piper, Dikic, and Lukacs, 2014). Please refer to Tab. 1.1 for more PTMs that are relevant to the work presented within the thesis. Common experimental techniques to identify and study PTMs are immunodetection with specific antibodies and mass spectrometric methods via stable isotope labeling by amino acids in cell culture (SILAC, Schmelzle and White (2006)).

Table 1.1: Reversible post-translational modifications (PTMs). A selection of reversible PTMs (post-translational modifications) relevant for the current study are listed below with biological examples.

PTM	**Modified residue**	**Exemplary function**
Phosphorylation	S, T, Y	activation of MAPK signaling components (Johnson and Lapadat, 2002)
Ubiquitination	K	regulation of receptor sorting and endocytosis (Piper, Dikic, and Lukacs, 2014)
SUMOylation	K	DNA damage repair, transcription regulation (Andreou and Tavernarakis, 2009)
Hydroxylation	P, K, N	hypoxia, regulation of interaction with other modifying enzymes ("crosstalk"), e. g. MAPK6, p53, Akt (Zurlo et al., 2016)
Nitrosylation	C	mediation of nitric oxid influence, e. g. on cardiac ion channels (Foster, Hess, and Stamler, 2009)
Lipidation	C	protein membrane anchor e. g. Ras GTPase (Nalivaeva and Turner, 2009)

1.2 Gene duplications as evolutionary playground

Gene duplications are often the starting point for the evolution of new functions. As stated earlier, orthologs more often preserve a function, while paralogs more likely gain a new function or expression profile. Nevertheless, the fate of a duplicated gene highly depends on the functional organization of the protein-coding gene, its function, expression profile, the duplication mechanism and of course its adaptive value for the organism in the current environment. Although, paralogs are an essential concept and seem to be very common, the vast majority of gene duplications have a negative or neutral effect on the organism's reproductive success (fitness) and thus will be purged from the genome.

1.2.1 Mechanisms of gene duplication

Duplications can arise by different mechanisms: whole genome duplication (WGD), tandem duplication and duplicative DNA or retro-transposition. The duplication mode may influence the selective pressure acting on the paralogs as discussed below. A WGD, namely the duplication of the entire genetic material of an organism, arises from polyploidy (van de Peer, Maere, and Meyer, 2009; van de Peer, Mizrachi, and Marchal, 2017). Autoploidy, the only known form of polyploidy in animals (Metazoa), is caused by incomplete division of the nuclei of a fertilized oocyte within the same species during cell division. Three major WGD events have shaped the evolution of deuterostomes, two WGDs at the base of vertebrates (1st Round WGD, 1R-WGD; Second Round WGD, 2R-WGD) and one WGD in the ancestor of teleosts (Third Round WGD, 3R-WGD, Meyer and van de Peer (2005)). These events led to an explosion of species and functional as well as morphological diversity leading to higher biological complexity (van de Peer, Maere, and Meyer, 2009). The term "2R-WGDs" will be used in the following to refer to the two rounds of WGD. The vertebrate 2R-WGDs allowed, among others, the diversification of the nervous and circulatory system (Roux, Liu, and Robinson-Rechavi, 2017), sensory organs such as the visual system and the development of paired appendages and body segmentation in jawed vertebrates. Hemoglobins (Hoffmann, Opazo, and Storz, 2010), opsins (Larhammar, Nordström, and Larsson, 2009) and the *Hox* genes (Garcia-Fernandez, 2005) are prominent examples of gene families that were retained after 2R-WGD.
A much smaller duplication, the tandem duplication, arises from recombination at non-homologous breakpoints during crossing over in cell division. This process results in one chromosome carrying a tandem duplication, while the homologous chromosome will have a deletion of the size of the duplicated fragment (Magadum et al., 2013). Tandem duplications are characterized by two identical genomic fragments situated next to each other on the same chromosome right after the event happened. Depending on the location of breakpoints, the duplicated genomic fragment might encompass several genes, one gene or parts of a gene.
The last mechanism discussed here is the duplicative DNA and RNA-based transposition (retrotransposition) that leads to the insertion of a genomic fragment at a random position within the genome. This mechanism stands in contrast to tandem duplications, where paralogs are at least initially situated in close proximity to each other. Some retrotransposable elements, such as long interspersed elements-1 (LINE-1) in mammals, recognize processed mRNAs and reverse transcribe these into complementary DNA (cDNA, Esnault, Maestre, and Heidmann (2000)). The insertion of these cDNAs into the genome is mediated by an endonuclease and ligase functionality that may be part of the retrotransposon (Kaessmann, Vinckenbosch, and Long, 2009). For that reason, young retrogenes usually have a poly-A tail and do not possess introns. Nevertheless, introns may be acquired over time and might increase expression of the retrogene (Fablet et al., 2009). The probability of retrotransposition increases with germline expression level (Zhang, Carriero, and Gerstein, 2004). While most retrogenes will degrade and vanish into the genomic background, others might be expressed and translated into a functional protein. Functionally expressed retrogenes usually have to acquire regulatory elements (promotors, enhancer etc., Kaessmann, Vinckenbosch, and Long (2009)). Parental genes represent, that possess a 5' upstream reading frame, represent an exception as they piggy-bag a promotor region upstream of a TSS within the retrogene. Another possibility for retrogenes to be immediately expressed is their insertion into an expressed region such as an open chromatin region or the intron of a transcribed gene (possibly leading to gene fusion, Marsh

and Teichmann (2010)). Due to their different genomic environment and thus likely different regulation as well as the lack of introns, retrogenes are not functionally equal to their parental genes at birth (Jun et al., 2009). They are thus more prone to develop a new functionality or expression pattern than their multi-exon paralog.

1.2.2 Evolutionary fate of duplicated genes - Should I stay or should I go?

Selection describes the propagation of a gene variant as consequence to its effect on the fitness of an individual (Vitti, Grossman, and Sabeti, 2013). If the gene variant increases the individual's fitness, it is propagated throughout the population (positive selection). On the other hand, specific variants might be disfavored leading to their elimination within the population and the maintenance of the original variant (negative or purifying selection). In the following, I will shortly describe the evolutionary pressure acting on protein-coding genes depending on the duplication mechanism that gave rise to the gene copy. Time-wise, two different phases are distinguished that follow the gene duplication event, which happens in the germline of a single organism. During the fixation phase, the frequency of the duplication variant increases within the population relative to the unduplicated variant. If the unduplicated variant vanishes, the duplicated variant is called fixed. The preservation phase starts, when the duplicated variant is fixed in the population (Innan and Kondrashov, 2010). Several models and selection pressures may apply to the same gene or gene family over time.
A gene duplication can be immediately advantageous for an organism, e. g. duplication of a transport gene that enables the organism to increase the uptake of a substrate (positive dosage, Lin and Li (2011)). This usually applies to genes that mediate an interaction between the organism and its environment, proteins needed in large quantities or are part of dosage-sensitive protein-protein interactions (Kondrashov et al., 2002). Following a selective sweep, the duplication variant will be fixed as it is highly advantageous for the organism. In the positive dosage scenario, both gene copies evolve under negative selection in the preservation phase. Positive selection on the duplication event also arises if the duplicated gene immediately gains a new, beneficial function, e. g. a retrogene that gains a new promotor and thus a different temporal and spatial expression profile (Vinckenbosch, Dupanloup, and Kaessmann (2006), section 1.2.1). This can also apply to the insertion of a newly duplicated gene into an existing reading frame possibly leading to the gain of a protein domain (Marsh and Teichmann, 2010).
WGDs are very rare events and usually a dead end in evolution (van de Peer, Mizrachi, and Marchal, 2017). This can be different under specific conditions, e. g. when ecological niches are not occupied during a mass extinction event that drastically changes the fitness landscape. In this case, a WGD can be an advantage as it mediates a higher vigour and might allow for faster adaptation of that species to the environment (van de Peer, Maere, and Meyer, 2009). It has been estimated that only about 10-20 % of the paralogs derived from an ancient WGD (ohnologs) are retained after fixation of the WGD in the population (Roux, Liu, and Robinson-Rechavi, 2017). While some ohnologs will be retained merely unchanged (under negative selection) to conserve the dosage ratio of interaction partners that might be duplicated, other ohnologs might evolve under positive selection enabling specialization and the gain of new functions (neofunctionalization) in the preservation phase. The first process might even facilitate the second in some cases (Thompson, Zakon, and Kirkpatrick, 2016). For other proteins, the majority of mutations will lead to misfolding and eventually aggregation of a toxic protein product (Yang et al., 2012). To counteract this

effect, negative selection will conserve the respective gene and has been postulated to maintain a high fraction of ohnologs expressed in the nervous system (Roux, Liu, and Robinson-Rechavi, 2017). Following along the line, WGDs have been connected to an increase of fraction of disease-connected genes (Singh et al., 2014).
Gene duplications with an immediate negative effect on the organism's fitness are not fixed within the population. There are plenty of examples of disease causing copy number variations in human, e. g. high copy number of *ERBB2* associated with aggressive forms of breast cancer (Peiro et al., 2004). In accordance with these observations, paralogs originating from duplications other than WGD, tend to show a retention pattern opposing to their ohnologs, i. e. with local duplications of highly expressed nervous system genes rarely fixed in populations (Roux, Liu, and Robinson-Rechavi, 2017) and local duplicates having a lower risk to be associated with disease (Singh et al., 2014).
If the duplication has neither an immediate negative nor a positive effect on the organism's fitness, selection on the duplicated copy is neutral during the fixation phase, i. e. the duplication variant is neither favored nor disfavored. The duplicated gene might get fixed in the population by genetic drift if the fixation phase is sufficient short. Both genes are subject to mutations that can lead to gain or loss of protein function (Innan and Kondrashov, 2010). The probability of a gain-of-function vs. a loss-of-function mutation highly depends on the function and structure of the individual protein (section 1.1.4). Loss-of-function is a much more likely consequence of a mutation (Behe, 2010). Deleterious loss-of-function mutations usually lead to the pseudogenization of one of the gene copies (Innan and Kondrashov, 2010). The gene copy might degrade and vanish into genomic background over time. The human genome encodes about 11,000 pseudogenes, of which about 7.5 % are transcribed (The ENCODE Consortium, 2012). Some of the transcribed pseudogenes might have gained a regulatory function as e. g. a ncRNA (Gulko et al., 2015).
In an alternative scenario, both copies accumulate loss-of-function mutations that lead to the degeneration of different functions in both copies (duplication-degeneration model, Force et al. (1999)). In this case, none of the two copies can fulfill the full functional repertoire of the former unduplicated gene. Both genes are specialized. In contrast to pseudogenization, purifying or positive selection acts independently on both copies in the preservation phase to maintain or exceed on the original function, respectively. A similar scenario applies to multifunctional genes after gene duplication that might be able to improve their original functions through gain-of-function mutations in both copies exceeding on the functional repertoire of the unduplicated gene (escape from adaptive conflict, Hughes (1994)). The oldest theoretical model on gene duplications is the neofunctionalization model proposed by Ohno (1970). This model assumes that both gene copies are exactly identical immediately after duplication, e. g. after a tandem duplication of the full-length gene. While one copy evolves under negative selection pressure in the preservation phase and maintains the original function, the other copy evolves under positive selection pressure. This copy might give rise to a different function.

1.3 Identification and annotation of protein homologs

Within this section, I will focus on current methods for the retrieval and annotation of homologous sequences given a protein family of interest and challenges in the field. For analysis of gene loss and gain patterns and the inference of possible functional changes for a protein family of interest, a complete set of annotations within the

genomes of interest is warranted. This goal can only be achieved, if the target genomes are exhaustively searched for encoded gene copies as public databases generated with automated gene annotation pipelines are frequently incomplete. Those automatic gene annotation tools usually do not consider fragmented gene loci. Challenges of current databases and gene annotation tools are discussed in the following with special focus on fragmented assemblies. This challenge has been a motivation for the development of the `ExonMatchSolver` pipeline (Chapter 2).

1.3.1 Challenges of available genome assemblies, gene annotations and sequence databases

Unfortunately, chromosomes that encode the genomic information of an individual cannot be sequenced as a whole. Instead, reads of about 150-300 bp length are usually generated during the genome sequencing process with an `Illumina` sequencer (Illumina Inc., 2017), while longer reads with a higher fraction of random errors recently became available with the development of the `PacBio` long-read sequencing technique (Eid et al., 2009). In order to reconstruct the original genomic sequence, reads are merged resulting in bigger genomic fragments in a process called genome assembly. A genome annotation is associated with a particular assembly version of the species' genome, which both might be refined over time (Eilbeck et al., 2009).
The computational and technical advances of recent years now enable single groups to successfully tackle such a genome project - a task only feasible to be completed by big consortia few years back (Lander et al., 2001; Waterston et al., 2002). Currently (as of October 2017), 372 deuterostome nuclear genomes are publicly available in the central database of the `NCBI` (O'Leary et al., 2016), that can potentially be mined to study a gene family of interest. The plurality of new genome releases is a valuable resource for reconstructing large-scale gene families, although at the same time it poses challenges for the bioinformatics community due to an increasing heterogeneity in available genome assemblies and annotations. The genome annotations and assemblies are usually submitted to public protein and nucleotide databases (such as `UniProt` (The UniProt consortium, 2015) or the `NCBI Genbank` (Sayers et al., 2012)) and are sometimes reanalyzed with unified workflows, e. g. the `Ensembl` (Cunningham et al., 2015) or the `NCBI` (Sayers et al., 2012) genome annotation pipelines. If manpower is available, manual curation polishes the gene annotations generated by automated methods, which are stored in "high-confidence", reviewed databases such as `RefSeq` or `Swiss-Prot`. Despite the best efforts of the biocurators community and continuing improvements, these data sources contain high levels of errors and inaccuracies (Carugo and Eisenhaber, 2010) that are virtually unavoidable given the volume of data that must be processed to create them.
Genome annotation is especially challenging for exotic genomes e. g. due to a high repeat content. Low sequencing coverage and low prior knowledge about gene models impose further hurdles (Yandell and Ence, 2012; Koepfli, Paten, and O'Brien, 2015). One particular difficulty is that most available genomes have unfinished assemblies, i.e., the corresponding genome assemblies consist of many, often short contigs and scaffolds, and genes span over more than one of these genomic fragments. Fragmentation is frequently not considered in the annotation process. The `Ensembl` pipeline, for instance, rejects matches covering less than 25 % of the query protein (Curwen et al., 2004). This results in missing, incomplete or inaccurate annotations of protein-coding genes, especially in assemblies with an average low contig length. Gene families might be affected differently with long genes and genes with micro-exons being especially challenging to annotate. Haug-Baltzell et al. (2015) and Horita et al.

(2012) have encountered these challenges on fragmented assemblies while annotating the dopamine receptor and the DUSP1 transcription factor families, respectively.
Assemblies built from PacBio reads are less fragmented in comparison to assemblies built from Illumina reads. Korlach et al. (2017) compared the gene annotations of two genomes that have been sequenced with both techniques, the genomes of humming bird and zebrafish. The study showed that the PacBio assemblies can resolve missing sequences in gaps and erroneous sequences adjacent to gaps and thus can contribute to the improvement of protein-coding gene annotations (Korlach et al., 2017). Although long-read sequencing techniques are becoming available to a broader community, the fragmentation of protein-coding genes across different genomic units is likely to persist in the near future. Within the Genbank database, 20.3 % of all eukaryotic genomes and 6.8 % of the animal genomes are at present assembled only to the contig-level (National Center for Biotechnology Information, 2017), while the vast majority of genomes is assembled to scaffold-level (66.8 % and 82.5 %, respectively). Even many of the genomes assembled to chromosomes still contain highly fragmented parts.
A particular difficulty of databases is the spread of erroneous annotations through similarity-based gene annotation methods, that use those erroneous sequences as queries. The errors propagate and the resulting erroneous annotations "poison" the experiment that make use of them (Yandell and Ence, 2012). The naming of genes in public resources adds another level of complication, and another potential source of error for the user, as nomenclature conventions are restricted to individual species or small groups of species. The HUGO Gene Nomenclature Committee is working to establish a coherent naming scheme for the genes in vertebrate genomes, aiming at a nomenclature that actually reflects homology as much as possible (Wain et al., 2002). In practice the retrieval of family members relies either on using databases of homologs such as Ensembl Compara (Vilella et al., 2009), OrthoDB (Waterhouse et al., 2013) and HomoloGene (Sayers et al., 2012), or on the use of similarity-based sequence search tools such as BLAST (Altschul et al., 1990). The use of public homology databases unavoidably is limited to the data included by its curators and restricted to the data sources, i.e., genome annotations, that have been selected for inclusion. Recently completed and still poorly annotated genomes are often not yet included.

1.3.2 Similarity search approaches for the identification of homologs without consideration of the gene structure

Homologs can be detected by a similarity search against databases or genomes. As databases are often incomplete or contain errors (section 1.3.1), mining of genomes can be necessary to retrieve a complete set of homologs of the protein family of interest. In this case, a short query sequence such as a protein-coding gene or protein sequence must be found within a long target sequence, the genome. If the query does not have an identical match in the target, a scoring matrix is applied to rank the possible solutions. The scoring matrix stores similarity scores between every character of the query and target alphabets. The scores for mismatches between query and target characters are position-independent and solely depend on the character identities. The described question corresponds to the local pairwise alignment problem. The Smith-Waterman algorithm calculates the best local alignment of query and target by employing dynamic programming, i. e. it finds the optimal solution of the local alignment problem given a specific scoring matrix and gap score.
Due to the enormous size of genomes and state-of-the-art databases, construction of all optimal alignments between query and target is not feasible. Commonly applied

methods such as BLAST rely on heuristics to speed up the search process. The BLAST algorithm first finds exact hits of query subsequences of length k (k-mer) in the target sequence. Those hits are extended until the score decreases and are then called high scoring segment pairs (HSPs, Altschul et al. (1990)). Overlapping HSPs above a score threshold are subsequently merged. For each hit, the algorithm returns a database-dependent expectation value (E-value) and a Smith-Waterman score, which is normalized on scoring matrix specific parameters and the search space (bit score). The bit score is thus independent of the applied scoring matrix. The E-value denotes the expected number of equally or better scoring hits to occur in this database. BLAST comes in different flavors applicable to different alphabet types of query and target (e. g. `tblastn` for querying an amino acid sequence against a nucleotide database).
An useful extension of the pairwise alignment is the multiple sequence alignment (MSA). In the following, the MSA will refer to an alignment of more than two full-length homologous sequences. Due to its complexity $O(n^l)$, which depends on the sequence length n and the number of sequences l, heuristics are applied for solving MSAs (for details on the algorithms, please refer to an introductory Bioinformatics book, e. g. Lesk (2014)). These heuristics usually sacrifice accuracy for speed. A common strategy is the progressive alignment, which adds sequences to the current alignment beginning with a sequence pair e. g. `Clustal Omega` (Sievers et al., 2011) or MUSCLE (Edgar, 2004). Other sequences are added progressively by following a pre-computed guide tree. The guide tree is a phylogenetic tree, that is often built based on pairwise sequence distances. In this case, the initial sequence pair is least distant (most similar). The specific heuristic implementations mainly differ in the generation of the guide trees and their computational representation.
The MSA reflects important properties of the respective homologous sequences such as the position-dependency of substitutions, insertions and deletions. For the specialized task of detecting a specific protein family of homologous sequences, it seems desirable to adapt the scoring matrices and gap costs to the specific sequence context and account for exactly those statistical properties. Eddy (1998) provides a framework for building probabilistic models, the profile Hidden Markov Models (pHMMs) from a MSA.
In a hidden Markov Model (HMM), observed states, e. g. different amino acids occurring in the same alignment column, are generated by a hidden process (Fig. 1.6). Allowed hidden states within the pHMM framework are the match, deletion, insertion as well as begin and end states. Those states are traversed in a linear fashion from profile start to end depending on transition probabilities between states (Eddy (2008), see direction of arrows in Fig. 1.6). Different extensions of this described, simple pHMM setup accommodate local alignments and multiple hits of the query profile against a target (Eddy, 2008). Apart from using these desirable probabilistic properties, pHMMs as implemented in the HMMER3 software, are based on a solid Bayesian framework calculating the posterior probability (pp) of the alignment ensembl. In addition to an optimal alignment score ("Viterbi score"), HMMER3 returns the Forward score, a log-odds likelihood score accounting for alignment uncertainty (Eq. 1.1). The null hypothesis, H_0, assumes that a target profile r is homologous to the query profile s, while the alternative hypothesis, H_1, states that r is unrelated to s (Eddy, 2009). Given the set of all possible alignments π of r to the query, the Forward score F is calculated as follows (Eq. 1.1, Eddy (2009)).

$$F = \log_2 \frac{\sum_\pi P(r, \pi | H_0)}{P(r|H_1)} \tag{1.1}$$

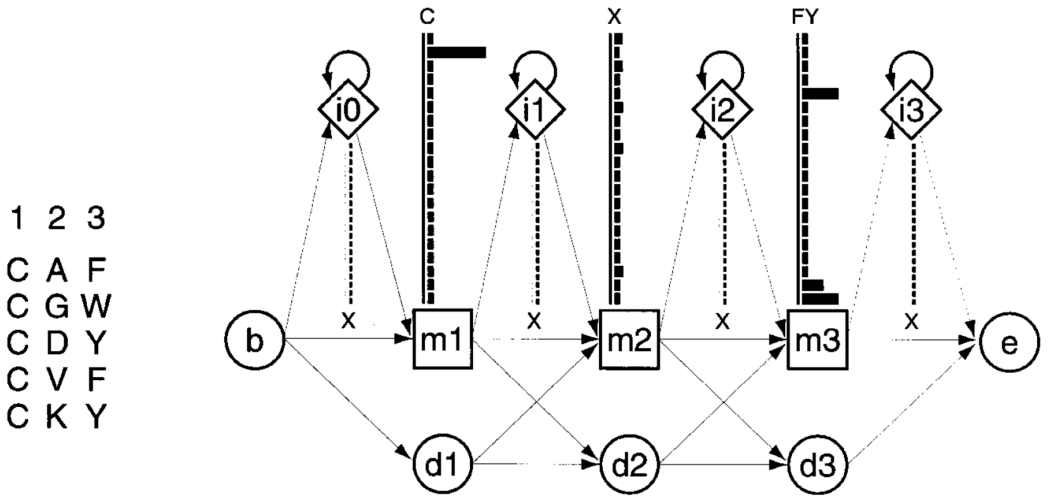

Figure 1.6: Schematic depiction of a simple profile hidden Markov Model. The shown profile hidden Markov model is built from a toy alignment (left) of five amino acid sequences. Every consensus column is depicted by a node, which can take a match state (square) or a deletion state (circle). Additional sequence in comparison to the consensus sequence is accommodated by the insertion states between the consensus alignment columns (diamond). The single states are traversed from left to right with specific transition probabilities between states (arrows). The match state and insertion state emit certain amino acids according to emission probabilities. Those are shown as amino acid frequency bars above the match and below the insertion state, respectively. The model is initialized at the begin state (b) and ends in the end state (e). The figure was taken from Eddy (1998).

`HMMER3` reaches a speed similar to `BLAST`, which is accomplished by several filtering steps of initial hits. As expected, the alignment of those profiles is often more accurate than the pairwise alignment of single sequences (e. g. by `BLAST`). Furthermore, pHMMs can detect more remote homologs making it a desirable framework for detecting complete sets of homologs (Eddy, 2011; Finn et al., 2015). Although the `HMMER3` suit offers numerous subprograms, e. g. `hmmbuild` for construction of a pHMM from a MSA and `hmmsearch` for searching databases with pHMMs, an equivalent of `tblastn`, the search of protein pHMM against a nucleotide database is not yet implemented (Eddy, Wheeler, and HMMER development team, 2015).

1.3.3 Gene structure aware gene annotation approaches applied to whole genomes

Due to the structure of a gene (section 1.1.1), detection of homologous or similar sequence stretches with means described in section 1.3.2 is not equal to gene annotation. A gene that encodes a functional protein, has a start and stop codon at the beginning and end of the open reading frame, respectively, maintains the codon reading frame in multi-exon transcripts and conserves SS patterns. Gene annotation approaches account for those properties to different extends. As annotation of a whole genome is a hard problem, state-of-the-art methods typically combine *ab initio* and similarity-based gene prediction methods. *Ab initio* gene prediction tools are usually trained on cDNA or RNA sequencing (RNA-Seq) data available for the species of interest to obtain probabilities to build (generalized) HMMs (`AUGUSTUS` (Stanke and Waack, 2003), `GENSCAN` (Burge and Karlin, 1997)), that model the statistical properties of transcribed genes. In contrast to pHMMs (section 1.3.2), where states roughly correspond

to alignment columns, states in the generalized HMM framework represent gene features (exon, intron, SS etc.) with specific properties (e. g. length, nucleotide and codon composition, see Carugo and Eisenhaber (2010) for a review). In general, *ab initio* methods massively overpredict open reading frames (Yandell and Ence, 2012), while the reliance on statistical properties causes them to frequently miss short exons consisting of few codons. *Ab initio* gene predictions can often be improved upon in detail, e. g. more exact gene boundaries, by similarity-based alignments.
The similarity-based methods benefit from available cDNA, expressed sequence tag (EST) or protein data from the same species (producing a *cis*-alignment) or from a closely related species (producing a *trans*-alignment, Brent (2008)). Key similarity-based methods in this context are spliced alignment algorithms; these align one single protein (profile) or cDNA/EST sequence at a time to a short genomic locus (`ProSplign` (Thibaud-Nissen et al., 2013), `Prot_map` (Softberry, 2007), `GeneWise` (Birney, 2000)) or to the whole genome (`exonerate -m est2genome` (Slater and Birney, 2005), `GenomeThreader` (Gremme et al., 2005)) while allowing for insertions in the target sequence (corresponding to introns) and considering SS patterns. First, gene loci are identified – either by an alignment heuristic (spliced aligners) or by a much faster mapping approach (e. g. `Spaln` (Gotoh, 2008) or `GMAP` (Wu and Watanabe, 2005)). Secondly, alignments are refined applying the exhaustive Smith-Waterman algorithm. One example of a spliced aligner is `exonerate` (Slater and Birney, 2005), the core of the `Ensembl` gene annotation pipeline. The implementation generates gaped alignments from HSPs and subsequently applies an extended version of the Smith-Waterman-Gotoh algorithm to a subset of alignments thereby accounting for introns, frame shifts and the translated amino acid sequence. For *trans*-alignments, the performance of all spliced alignement tools highly depends on the distance of query and target species.
Another known difficulty for similarity-based methods is the identification of tandemly duplicated genes (Thibaud-Nissen et al., 2013). `Splign/ProSplign`, part of the `NCBI` gene annotation pipeline, explicitly tackles this problem by implementing a dynamic programming solution. Gene loci are identified by finding chains of valid HSPs that form non-overlapping compartments, while the score over all compartments is maximized (see Thibaud-Nissen et al. (2013) for details). In comparison to other spliced aligners, `Splign` is good at finding small exons (Kapustin et al., 2008).
As mentioned above, a combination of *ab initio* and similarity-based methods is implemented in the most popular and widely used gene annotation pipelines, i. e. `Ensembl` (Curwen et al., 2004), the `NCBI` eukaryotic genome annotation pipeline (Thibaud-Nissen et al., 2013) and `AUGUSTUS` (Keller et al., 2011). Pipelines typically differ in their pre- and post-processing steps as well as how specific tools are combined. Specifically, an extension of the *ab initio* gene prediction tool `AUGUSTUS` improves on prediction accuracy by giving a score bonus for the predicted transcript if a pHMM block from a known protein matches (Keller et al., 2011). Native to similarity-based methods, Iwata and Gotoh (2012) improved exon accuracy of their spliced aligner `Spaln2` in plants and fungi by incorporation of branch point signals and oligomer composition. All combiners usually constitute a trade off between accuracy (from similarity-based methods) and speed (from *ab initio* methods). Although steps in gene annotation pipelines have been optimized - some even for years - it is thus not surprising that they still may make mistakes such as over- and under-predicting small introns and exons. Even extensive EST or RNA-Seq data sets may be incomplete. Both false positive and false negative predictions are propagated by the similarity-based methods and can only be rectified, in part, by the diligent work of human curators (section 1.3.1).

As opposed to *ab initio* and similarity-based methods that are applied to single genomes, a third group of gene annotation strategies is based on the identification of regions that are conserved across genomes of different species implying a conservation of function. Conservation-based methods have been employed for the annotation of nc elements, i. e. regulatory regions or ncRNAs (Nitsche et al., 2015), as well as protein-coding genes (Washietl et al., 2011; Sharma, Schwede, and Hiller, 2017). Conserved regions are identified from whole genome alignments or alignments of homologous genomic regions. Although whole genome alignment methods are improved and updated to increase accuracy and sensitivity (Sharma, Schwede, and Hiller, 2017; Suarez et al., 2017), they frequently have difficulties with resolving paralogy and orthology.

1.3.4 Graph-based inference of orthology relationships on proteomes

The resolution of the orthology and paralogy relationships is important to draw conclusions about sequence-function relationships as orthologs have often the same or similar functions, while the function of paralogs frequently differs (section 1.1.3). Graph-based methods are employed to resolve the orthology relations of all proteins (the proteome) encoded in two or more genomes. They implement two key steps, the retrieval of pairwise similarity scores, and clustering of the hits to resolve the orthology relations. Most methods assume that 1:1 orthologs score best to each other (i. e. are bidirectional best hits). Similarity is frequently measured by the `BLAST` bit score (Li, Stoeckert, and Roos, 2003; O'Brien, Remm, and Sonnhammer, 2005; Lechner et al., 2011) or the Smith-Waterman score (Waterhouse et al., 2013). The scores are used as edge weights connecting proteins (nodes). The nodes are usually clustered under relaxation of the reciprocal best hit requirement, e. g. by `OMA` (Train et al., 2017). `OrthoDB`, for instance, clusters hits progressively by extending the pairwise best hits to best hit groups of three on different clade levels.
As relying on all-against-all best hits of the input proteomes, the performance of graph-based approaches highly depends on the quality and completeness of the input data. Most algorithms rely on the common protein databases listed in section 1.3.1 and additionally consider specialized genome annotation databases such as `FlyBase` (McQuilton et al., 2012). The precision of orthology group assignments is increased by consistency checks of the input protein sequences, stringent similarity search and clustering settings as implemented in `OMA`. The higher precision often comes at expense of recall (Altenhoff et al., 2016). This leads to unwanted effects; genes that are missing from the annotation or excluded during the consistency checks are inferred as gene losses, while fragmented genes might appear as false positive group assignments (Train et al., 2017).

1.3.5 Gene annotation and inference of orthology relations on fragmented assemblies - chance and challenge

As pointed out in sections 1.3.1 and 1.3.4, genes spanning several genomic fragments in fragmented assemblies cause problems during gene annotation and orthology assignment. On the other hand, they provide the chance to improve the current genome assembly. In fact, a specialized branch of genome assemblers makes use of gene annotations and external information such as physical maps to scaffold fragmented genomes, e. g. `PEP_scaffolder` (Zhu et al., 2016), `GPM` (Zhang et al., 2016), `Swips` (Li and Copley, 2013), `ESPRIT` (Dessimoz et al., 2011). The scope of

those assembly strategies is usually to deliver a high quality assembly of a single genome rather than the optimization of annotations of specific gene families. One of the early implemented approaches, ESPRIT (Dessimoz et al., 2011), first annotates genes with AUGUSTUS. Next, orthologs between the genome of interest and several reference genomes are inferred with the OMA algorithm. Two fragments from the fragmented genome are scaffolded if they consistently map to a single gene in different reference genomes. As discussed by Dessimoz et al. (2011), the approach's weakness lies in the resolution of close, but distinct paralogs, e. g. in genomes of tetrapods and teleosts (section 1.2.2). This problem is tackled by the time-expensive Swips pipeline (Li and Copley, 2013), that first locates gene loci by a tblastn similarity search and refines the gene structure with GeneWise. In-paralogs are considered by an α parameter, that determines the strength of coupling of a protein sequence and its hits on the contigs of the fragmented assembly.
Especially in studies that focus on single gene families, completeness of the gene annotation with respect to coverage and resolution of the orthology relationships is critical (section 1.1.3). The latter is rarely considered in the context of gene annotation on fragmented assemblies. For example, in the SGP2 framework (Parra et al., 2003), *ab initio* gene prediction (geneid) and similarity search (tblastx) are combined. SGP2 assumes that hits on different fragments originate from a non-assembled shotgun genome and will summarize these hits to one gene prediction by re-scoring of HSPs. Thus, different, highly similar paralogs tend to be merged into a single gene prediction. The combined mapping/alignment tool GMAP (Wu and Watanabe, 2005), which was originally intended to uncover chimeric ESTs, maps cDNA/ESTs to multiple genomic loci. This method theoretically allows for annotation of genes in a fragmented genome, although to my knowledge application of GMAP has been limited to *cis*-alignments (Wu and Watanabe, 2005). The Scipio system (Keller et al., 2008) was developed originally for *cis*-alignments of proteins and cDNAs and later has been extended to *trans*-alignments (Hatje et al., 2011). It proceeds stepwise: (1) Blat alignment; (2) Gap closing in the query sequence to detect short exons using a Needleman-Wunsch alignment; (3) Assembly of the blat hits; and (4) Intron border refinement (Hatje et al., 2011). Recent refinements to accommodate the needs of particular query genes are described in Hatje and Kollmar (2011), Pillmann et al. (2011), and Hammesfahr et al. (2015).
The problem of assembling genes from multiple genomic fragments becomes particularly difficult in cases where multiple close paralogs are present. A frequent error is the construction of chimeric gene models that thread through fragments belonging to different paralogs, see e. g. Pavesi et al. (2008).

1.4 Methods that benefit from more complete gene annotations and exact orthology relationships

Numerous methods built on gene annotations or MSAs of orthologs. They thus benefit from high-quality and with regard to paralog number and sequence coverage more complete annotations. Among them are methods for phylogenetic inference, natural selection analysis and inference of specificity determining positions (SDPs), covered in this section. They help to infer information about the sequence-function relationship as demonstrated in this thesis when studying the evolution of the arrestin protein family (Chapter 3).

1.4.1 Phylogenetic inference in a nutshell

There are two classes of methods to infer phylogenetic trees from alignments, the distance-based (semi-parametric) and character-based (parametric) methods. Distance-based methods condense character-based similarity by calculating pairwise distances between all sequences and subsequently reconstruct the tree based on the distance matrix, e. g. neighbor joining. In contrast, character-based methods such as Maximum Likelihood (ML) or Bayesian tree inference, model evolution explicitly by examining character or alignment columns (e. g. nucleotide, codon, amino acid) of all sequences assuming that all characters evolve independently. The effects and differences of phylogenetic tree reconstruction based on nucleotide, amino acid or codon MSAs are not well understood. Nucleotide MSAs usually harbor more information than amino acid MSAs, which can be beneficial for resolving phylogenetic relationships of very close sequences with few changes. On the other hand, amino acid MSAs represent homology often better for distant proteins as the underlying nucleotide sequence might have undergone multiple substitutions. Amino acid and nucleotide substitution models represent different approximations of the underlying evolutionary process, none of which is able to describe this process in its entity. Alignment insecurity further contributes to the inaccuracy of the modeled process. As this section very briefly introduces the underlying principles, please refer to Page and Holmes (2005), Gascuel and Steel (2007) and Drummond and Bouckaert (2015) for a more thorough explanation on substitution models and the ML method as well as Bayesian inference in phylogenetics, respectively.

Nucleotide, amino acid and codon substitution models

One important aspect of parametric models that is used to describe the evolutionary process are substitution models. Among other aspects are the molecular clock, rate heterogeneity and the tree generation process. Substitution models differ in the rates assumed for character conversions (in the following referred to as transitions unless explicitly specified otherwise). The substitution process is modeled as a continuous-time, Poisson-distributed Markov chain, where every character evolves independently. The probability of observing s after a transition from r to s has happened at time point t, depends on the instantaneous transition rate Q and the time point t (Eq. 1.2). The Markov process is furthermore assumed to be time-reversible ($P_{rs}(t) = P_{sr}(t)$) and stationary.

$$P_{rs}(Q,t) = e^{Qt}; r \neq s \tag{1.2}$$

Nucleotide substitution models describe the transition probabilities for conversion of every of the four nucleotides into each other ($r, s \in A, G, C, T$, Eq. 1.3). The simplest model, Jukes Cantor (JC69, Jukes and Cantor (2013)), assumes constant rates for all possible substitutions and equal probabilities (frequencies) of the four nucleotides. For JC69, the transition probability of nucleotide r to nucleotide s can be obtained analytically with one free parameter (transition rate α) estimated from the data (Eq. 1.3).

$$P_{rs}(\alpha,t) = \begin{cases} \frac{1}{4}(1-e^{-4\alpha t}) & \text{if } r \neq s \\ \frac{1}{4}(1-e^{-4\alpha t}) + e^{-4\alpha t} & \text{if } r = s \end{cases} \tag{1.3}$$

This model is not suitable for reconstruction of sequence evolution over long evolutionary distances. For this kind of application, more complex models such as

the Hasegawa, Kishino and Yano 1985 model (HKY85) with five free parameters represent better approximations. It accommodates two different rates for the two possible nucleotide substitution types, nucleotide transitions (G $\leftrightarrow$ A, T $\leftrightarrow$ C) and nucleotide transversions (C $\leftrightarrow$ A, T $\leftrightarrow$ G, T $\leftrightarrow$ A, C $\leftrightarrow$ G). Moreover, HKY85 allows for different frequencies of the four nucleotides (Hasegawa, Kishino, and Yano, 1985). The most general time-reversible nucleotide substitution model (GTR) concedes six different substitution rate parameters (for all possible $r \leftrightarrow s$ combinations under assumption of time reversibility) and different frequencies of the four nucleotides (nine free parameters in total).
The substitution rate across a gene or protein sequence is usually not uniform, but heterogeneous with some positions being constant e. g. the catalytic core, while other positions greatly vary across the alignment e. g. in loop regions of a protein (Echave, Spielman, and Wilke, 2016). The variation of substitution rates across a protein (rate heterogeneity) can be incorporated into the so-called mixture substitution model and are denoted by capital letters added to the substitution model abbreviation (+G, +I). Mixture models allow for rate heterogeneity following a γ-distribution with a fixed number of different rates (discrete γ-model, + G) and can accommodate a fraction of invariable sites (+I). The γ-distribution is described by the shape parameter, which specifies its mean with a fixed scale parameter. The CAT model accommodates rate heterogeneity be using a single rate for every site, which requires less memory than the γ-model and can thus be applied to very big data sets (Stamatakis, 2006).
In contrast to simple nucleotide substitution models, transition probabilities for complex substitution models such as GTR or amino acid substitution models do not have a closed analytical form. The transition probabilities are inferred by complex mathematical operations from the Q matrix, which stores instantaneous (time-independent) transition rates q_{rs} (please refer to Kosiol and Goldman (2005) for details on the inference of the Q matrix). For amino acid substitution models, the instantaneous rates (off-diagonal entries of the Q matrix) are inferred from empirical amino acid replacement matrices such as the Point Accepted Mutation matrix (PAM, Dayhoff, Schwartz, and Orcutt BC (1978)) or the Blocks Substitution Matrix (BLOSUM, Henikoff and Henikoff (1992)). Those amino acid replacement matrices are also used as scoring matrices for amino acid alignments (section 1.3.2). Early amino acid replacement matrices such as PAM and Jones-Taylor-Thornton (JTT, Jones, Taylor, and Thornton (1992)) were derived from highly similar pairwise protein alignments (PAM250: identity $>$ 85 %) by normalizing the substitution counts by the pairwise protein divergence. As those matrices do not account for multiple substitutions, their application is limited to protein sequences with a similar and low level of divergence. More recently, Whelan and Goldman (2001) used a ML approach to obtain the Whelan and Goldman (WAG) amino acid replacement matrix from a multiple alignment of globular proteins. The Le and Gascuel (LG) matrix extends on this approach allowing for rate heterogeneity during the ML estimation on a much bigger and diverse MSA (Le and Gascuel, 2008). Apart from those general models, plenty of amino acid replacement matrices are based and targeted on a specific class of proteins e. g. the G protein-coupled receptor (GPCR) membrane domain (Rios et al., 2015) or viral proteins (Dang et al., 2010).
Inference of natural selection acting on protein-coding genes of different species is based on counting the substitutions that change or do not change the amino acid identity (non-synonymous and synonymous substitutions, respectively). The calculation of the ratio of synonymous and non-synonymous mutations requires the modeling of codon substitutions. The codon model of Goldman and Young, implemented in `codeml` (Goldman and Yang, 1994), takes into account the nucleotide

transition/transversion ratio κ and the frequency p_s. The instantaneous rate for transition of codon r to codon s is defined as q_{rs} (Eq. 1.4). ω measures the selective pressure.

$$q_{rs} = \begin{cases} 0, & \text{if } r \text{ and } s \text{ differ at more than one codon position} \\ p_s, & \text{if } r \text{ and } s \text{ differ by a synonymous transversion} \\ \kappa p_s, & \text{if } r \text{ and } s \text{ differ by a synonymous transition} \\ \omega p_s, & \text{if } r \text{ and } s \text{ differ by a non-synonymous transversion} \\ \kappa\omega p_s, & \text{if } r \text{ and } s \text{ differ by a non-synonymous transition} \end{cases} \tag{1.4}$$

As apparent from Eq. 1.4, the model considers only codon positions with at most one substitution. Similar to the JC69 and HKY85 nucleotide substitution models, different codon models exist that consider differences in codon or nucleotide frequencies to a different extent and thus differ in the calculation of p_s. The simple F3X4 model assumes different target nucleotide frequencies for every of the three codon positions, while the F61 model assumes different frequencies for every of the 61 amino acid encoding codons. The F61 model is thus applicable to data that is subject to codon bias, the deviation of the codon frequency of a specific gene or species from the expected, equal codon frequencies.

Maximum likelihood tree inference

ML seeks the topology and branch length of the tree that maximizes the probability of observing the given data under a specific evolutionary model M. The data D is given as a MSA of length n consisting of columns $x_1, ..., x_n$. Some parameters such as the character frequencies are often estimated from the data directly and kept fixed. All other parameters of the ML tree are chosen such that the likelihood (L) of the data is maximized (Eq. 1.5).

$$L(M; x_1, ..., x_n) = P(D|M) = P(x_1|M) \cdot ... \cdot P(x_n|M) = \sum_{i=1}^{n} \log(P(x_i|M)) \tag{1.5}$$

Calculation of L encompasses the calculation of the probability of observing every single MSA column ($P(x_i|M)$) for a fixed set of parameters under a specific evolutionary model M. The probabilities for unknown, ancestral character states at inner nodes are summed up over all possible character states for every node. The results of ML inference are thus highly influenced by the model assumptions made, e. g. choice of substitution model, shape parameter of rate heterogeneity.
ML inference is implemented so that L is initially calculated for a tree given a starting set of parameters θ (topology, tree length etc.) and subsequently compared to the L of trees with different parameter sets. The parameter set is successively varied which corresponds to the exploration of tree space. As ML inference of phylogenetic trees is NP-hard (Chor and Tuller, 2005), different heuristics have been implemented to approach good tree solutions. The space of possible tree topologies can be very big depending on the number of leaves l ($\frac{(2l)!}{l!(l+1)!}$) and contain local optima.
During my thesis, I apply `PhyML` (Guindon et al., 2010; Guindon, Gascuel, and Rannala, 2003) for building trees from amino acid MSA with several hundred leaves. `PhyML` generates a starting tree by neighbor joining, which is subsequently improved by exploring the tree space by two heuristics, subtree pruning and regrafting (SPR) and nearest neighbor interchange (NNI) (Hordijk and Gascuel, 2005; Guindon et

al., 2010). During SPR, a subtree is joined to another branch of the previous tree ("regrafted"), which allows a faster movement through tree space in comparison to NNI, where only nearest neighboring subtrees are exchanged. For very big amino acid MSAs with several thousand leaves, I apply the approximately-best-ML method `FastTree`, which combines NNI, SPR and additional heuristics. The implemented heuristics lead to a less efficient exploration of tree space by reduction of optimization and SPR steps. `FastTree` furthermore implements an approximation of the discrete rate heterogeneity model, the CAT model (Price, Dehal, and Arkin, 2010). The resulting speed-up comes as a trade-off with a lower number of correct splits in comparison to the two most popular ML inference programs, `PhyML3.0` and `RaxML` (Price, Dehal, and Arkin, 2010).
While the systematic error in tree inference caused by violations of different model assumptions is difficult to measure, the sampling error can be assessed by nonparametric bootstrapping. During bootstrapping, alignment columns are sampled n times with replacement from the original alignment (Felsenstein, 1983). Individual ML trees are constructed from the resulting n alignments. The relative proportion of specific splits (bootstrap support, BS) reflects the stability of that split during re-sampling. They are thus taken as a conservative surrogate for "confidence intervals" in the ML inference (Efron, Halloran, and Holmes, 1996).
As discussed above, different models exist that seek to approximate the processes which generated the phylogenetic tree. The Likelihood ratio test (LRT) and information criteria such as the Akaike Information Criterion (AIC, Eq. 1.6, Akaike (1974)) or the Bayesian Information Criterion (BIC, Eq. 1.7, Schwarz (1978)) are approaches to compare the fit of different models (models 0 and 1) to the data in the ML framework. The goal of AIC and BIC is to evaluate the trade-off between gain in likelihood vs. increase in model parameters (degree of freedom, K, Eq. 1.8) and thus to prevent potential overfitting.

$$AIC = -2\ln(L) + 2K \tag{1.6}$$

$$BIC = -2\ln(L) + K\ln(n) \tag{1.7}$$

with n - sample size, L - Maximum Likelihood under model M.

$$K(0,1) = \theta_1 - \theta_0 \tag{1.8}$$

with θ being the total number of model parameters. Both, BIC and AIC, are estimators of the Kullback–Leibler divergence between the tested and the unknown true model. BIC penalizes the number of free parameters more than AIC.
Model selection for nucleotide and amino acid substitution models based on BIC and AIC are implemented in the commonly used programs `JModelTest` (Posada, 2008; Darriba et al., 2012) and `ProtTest` (Abascal, Zardoya, and Posada, 2005; Darriba et al., 2011), respectively.
The LRT compares two nested hypothesis, the null hypothesis (H_0) and the alternative hypothesis (H_1), whereby nested means that H_1 can be simplified to H_0 (Eq. 1.9).

$$2\Delta L = 2(\ln(L_1) - \ln(L_0)) \tag{1.9}$$

For obtaining an empirical P-value of how likely H_0 is falsely rejected, the probability distribution of the test statistic $2\Delta L$ can be approximated by a χ^2 distribution with K degrees of freedom, if none of the H_1 parameters is fixed at the boundary of H_0 (Wilks, 1938).

Bayesian tree inference

The key idea of Bayesian inference is the assumption that all model parameters originate from a probability distribution and are not fixed values as assumed in ML. The goal of Bayesian inference is to characterize the pp distribution of the MSA given a set of (continuous) model parameters ($P(M|D)$). The resulting posterior probabilities are easy to interpret in contrast to BS values as they denote the actual probability of seeing this specific parameter marginalized over all possible model parameters. A credibility interval can be reported by e. g. specifying 95 % of the pp density. Bayesian inference is based on the Bayes Theorem:

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)} = \frac{P(M)P(D|M)}{\int_M P(D|M)P(M)} \tag{1.10}$$

with $P(M|D)$ - pp, $P(M)$ - prior probability, $P(D|M)$ - likelihood.
The denominator of Eq. 1.10, the probability of the data, is a normalization factor that ensures that the joint probabilities of the data and all possible model parameters ($\int_M P(D|M)P(M)$) sum up to one. It is also called the marginal likelihood, as $P(D)$ marginalizes over all possible model parameters. The pp landscape is very complex and multi-dimensional as depending on (continuous) model parameters. As the exact calculation of $P(D)$ is thus computationally expensive, another heuristic is applied to get an estimate of the pp.
The Markov Chain Monte Carlo algorithm (MCMC, Hastings (1970) and Metropolis et al. (1953)) is commonly applied to explore the pp landscape in Bayesian tree inference. The algorithm samples model parameters proportionally to the actual pp. The MCMC according to Metropolis et al. (1953) with a symmetric proposal distribution works as follows: (1) Initialization at a random starting point x_0 in the probability landscape; (2) Iteration composed of (a) change of the position to x' ("step") based on a proposal distribution ($q(x'|x)$); (b) calculation of an acceptance rate of this new position (Eq. 1.11, 1.12); (c) Acceptance of the new point x' if $\alpha > 1$, otherwise drawing of a number n from a uniform distribution [0, 1] and acceptance if $n < \alpha$.

$$\frac{\frac{P(D|M')f(M')}{P(D)}}{\frac{P(D|M)P(M)}{P(D)}} = \frac{P(M'|D)}{P(M|D)} \tag{1.11}$$

$$\alpha = \begin{cases} \frac{P(M'|D)}{P(M|D)} \frac{q(x|x')}{q(x'|x)} & \frac{q(x|x')}{q(x'|x)} = 1 \text{ (symmetric)} \\ \frac{P(M'|D)}{P(M|D)} \frac{q(x|x')}{q(x'|x)} & \text{if } q(x'|x) \neq q(x|x') \text{ (asymmetric)} \end{cases} \tag{1.12}$$

The parameters (pp, branch lengths etc.) are saved every few hundred or thousand iterations (called generations). As the parameters of consecutive steps are auto-correlated due to a usually small step size (change of only one or few parameters in every step), not every sample is considered.
The calculation of the acceptance rate α is comparably easy as the marginalized likelihood $P(D)$ does not have to be calculated and is canceled out from the equation (Eq. 1.12). Hastings (1970) proposed a modifying constant, the Hastings ratio (q), to correct the acceptance ratio for asymmetric proposal distributions (Eq. 1.12). The efficient exploration of the parameter space is subject to ongoing research (Heled and Drummond, 2008; Höhna, Defoin-Platel, and Drummond, 2008; Wu, Suchard, and Drummond, 2013).
Prior knowledge about the evolutionary process that generated the data such as the molecular clock acting, existing monophyletic groups, phylo-geography etc. can be

incorporated by specifying distributions or bounds on model parameters (Drummond and Bouckaert, 2015). In my thesis, I am concerned with gene trees. Those are generated by a birth–death process caused by speciations, duplications and gene loss events (Zhao et al., 2015). As the gene tree encompasses species across all deuterostomes with very different population sizes, generation times and selection pressures, I decided to use the relaxed molecular clock model. This clock model allows for variation of the substitution rate across branches of the tree. Substitution models were chosen by model testing in the ML context (known as empirical Bayes). In order to avoid over-powerful priors on parameters, where no apparent information was available, those priors were chosen to be non-informative (diffuse). Moreover, different priors were tried to exclude confounding effects.
In the Bayesian framework, models are compared and selected based on evaluation of the Bayes Factor (BF), which measures the fit of the models to the provided data. The BF is the ratio of the marginal likelihoods of the two models under comparison (here model A and model B, Eq. 1.13, Jeffreys (1935)). In practice, the $\log(BF)$ is compared with values >1 and >3 indicating strong and very strong support for model A, respectively (Kass and Raftery, 1995).

$$BF(A,B) = \frac{P_A(D)}{P_B(D)} \tag{1.13}$$

Path sampling (Lartillot and Philippe, 2006) and the stepping stone method (Xie et al., 2011) outperform the harmonic mean estimator for estimating the marginal likelihood in selection of relaxed molecular clock and demographic change models as shown by Baele et al. (2012) and Baele and Lemey (2013). Path sampling and the stepping stone method rely on drawing samples from a number of different distributions along the path from the prior ($\beta = 0$) to the posterior distribution ($\beta = 1$, Eq. 1.14). Those paths differ in their power posterior β [0,1]. For a specific model M, the path q is defined as the product of the likelihood function and the prior (Eq. 1.14).

$$q^{\beta}(\theta) = P(D|M,\theta)^{\beta} P(\theta|M) \tag{1.14}$$

with θ - model parameters.
Both methods differ in the way how the marginal likelihood is estimated from the empirical likelihoods drawn along the path (see Lartillot and Philippe (2006) and Xie et al. (2011) for details). Practically, the choice of models is further limited to those, for which different runs converge to the same parameter estimates after a computationally reasonable number of generations.

Phylogenetic tree based inference of orthology relationships

Phylogenetic tree-based orthology predictions can deliver different information on single gene families as compared to purely graph-based orthology inference methods as described in section 1.3.4 (Altenhoff and Dessimoz, 2012). The orthology assignment relies on the mapping of speciation, duplication and loss events on the gene tree given a species tree. This process is called tree reconciliation. Most phylogenetic tree-based methods minimize the number of loss and duplication events according to the parsimony principle or explicitly model the birth–death process in a Bayesian framework (Akerborg et al., 2009). In a complex scenario with missing data caused by gene loss, missing gene annotations or insufficient species sampling (Fig. 1.7), tree reconciliation under minimization of duplication and loss events might not resolve the correct homology relationship (Zallot et al., 2016). For example, B2 and C3 appear

as 1:1 orthologs in the gene tree on the right hand-side in Fig. 1.7, although they are out-paralogs (left). Another difficulty for inferring the homology relationship even without duplications can be horizontal gene transfer (transfer of a gene from species A to species B), gene conversion, recombination, gene flow/migration and incomplete lineage sorting. It is widely accepted that recent horizontal gene transfer is not a common phenomena in vertebrates (Crisp et al., 2015). Nevertheless, gene flow and incomplete lineage sorting between closely related species or populations can result in a similar effect in vertebrates: individual gene trees do not correspond to the species tree (Siepel, 2009). Recombination and gene conversion between paralogs within the same species can further create different tree topologies for different parts of the same gene (Cortesi et al., 2015). The reconciliation problem gets even more difficult if the gene and species tree are not fully resolved or insecurities in tree inference are considered (see Altenhoff and Dessimoz (2012) for a review). For reconciliation of the arrestin gene and the deuterostome species tree according to the parsimony principle, no program was applied as the gene family is rather small with four homologs per species.

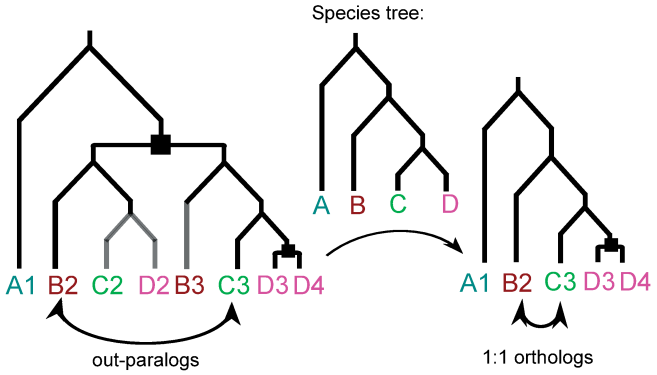

Figure 1.7: Missing data can lead to misassignment of paralogs and orthologs. The shown gene experienced two gene duplications in the given evolutionary time frame (duplications shown as squares, left). Assuming missing data or loss of the homologs in the grey branches, a different gene tree is inferred under the maximum parsimony scenario given the species tree (speciations shown as slopes, middle). In this different gene tree (right), out-paralogs seem to be 1:1 orthologs.

1.4.2 Inference of natural selection in inter-species data sets

Although selection is working on populations, methods exist that can be applied to detect selection between sequences of different species (section 1.2.2). For a protein-coding gene, these substitutions can either change the amino acid sequence – called non-synonymous substitution – or change the codon used, but not the amino acid encoded by this codon, referred to as synonymous substitution (section 1.4.1). The ratio of substitution rates of synonymous (dS) and non-synonymous changes (dN) ($\omega = \frac{dN}{dS}$) is a measure of the selective pressure acting on a protein-coding gene (Bielawski and Yang, 2005). Amino acid properties are not considered in $\frac{dN}{dS}$ models as the models assume that every non-synonymous change increases the fitness of the organism regardless of their physico-chemical properties or restrictions coming from the amino acid arrangement within the protein structure. Most protein-coding genes are conserved and thus evolve under strong purifying (negative) selection with a much higher rate of synonymous changes than non-synonymous changes ($\omega << 1$) (Eirin-Lopez et al., 2004). Pseudogenes evolve under neutral selection with equal

rates of synonymous and non-synonymous changes ($\omega = 1$) as selective pressure is relieved. These genes thus vanish into the genomic background over time (Gilad et al., 2003). A well studied example of a protein family that evolves under positive selection ($\omega > 1$) are the rare allelic variants of the major histocompability complex. Positive selection on the major histocompability complex drives high diversity of the cell surface protein that interacts with pathogens (Ejsmond and Radwan, 2015).
In order to infer selection based on $\frac{dN}{dS}$ methods, protein-coding genes have to be aligned with an amino acid-aware nucleotide aligner such as `MACSE` (Ranwez et al., 2011). Amino acid-aware nucleotide aligner first translate the nucleotide sequence into an amino acid sequence that is aligned and converted back into a nucleotide sequence. The `MACSE` algorithm optimizes the pairwise alignment score based on the amino acid replacement matrix while considering stops and frameshifts caused by deletions. As other MSA heuristics, `MACSE` follows a progressive strategy, where alignments are aligned to each other starting from the leaves of a guide tree (Ranwez et al. (2011), section 1.3.2). Simple models for the detection of natural selection based on $\frac{dN}{dS}$ assume a constant ω rate across all sites of a protein and all branches of a phylogenetic tree, which is highly unrealistic as discussed multiple times. More sophisticated models (site-models, branch-site models for natural selection) are implemented in a ML framework in the program `codeml`, part of the `PAML` package. They allow for a statistical distribution of positions across different site classes z with different ω_z rates (Yang and Nielsen, 2002; Zhang, Nielsen, and Yang, 2005).
The branch-site test implements a biological scenario, where selection acts on only a fraction of sites over a limited period of time, e. g. after a gene duplication or to accommodate changes in the environment (episodic evolution). In comparison to others tests, the branch-site model is especially sensitive. In the branch-site model, a foreground branch (FG) is defined with some sites evolving under positive selection, while the same sites evolve under purifying or neutral selection in the background branch (BG, Tab. 1.2). The models compared in the branch-site test differ in one free parameter (Model A H_1: 5, H_0: 4 parameters). For model comparison with the LRT (section 1.4.1), $2\Delta L$ should be compared to a 50:50 mixed distribution of χ^2 and 0 (Zhang, Nielsen, and Yang, 2005). Comparison to a χ^2 distribution is common practice and renders the test and empirical P-value more conservative (Zhang, Nielsen, and Yang, 2005). Furthermore, Yang and Nielsen (2002) strongly emphasize that branches to be tested for positive selection should be selected based on some prior biological knowledge to avoid false positives. In the current work, this *a priori* knowledge is given by gene duplication events. I test the branches immediately following the gene duplication under the assumption that the newly formed paralogs likely acquired a new function on this branch (section 1.2.2).
The pp for exact positions to be under positive selection can be calculated by application of the Bayes Theorem (Eq. 1.10). Proportions of positions (p_z) falling into the different site classes z estimated during ML inference are used as prior (empirical Bayes, Eq. 1.15).

$$P(z|x_i) = \frac{P(x_i|z)p_z}{P(x_i)} \tag{1.15}$$

with z - site class and i - position.
The pp of a position x_i belonging to one of the site classes z (with a fixed, prior ω_z) is corrected for ML estimation errors via uniform priors of some parameters (ω_0, ω_1 in the branch-site model) in the Bayes Empirical Bayes (BEB) method (Yang, Wong, and Nielsen, 2005). The BEB method has a much lower false positive rate (10 % with 90 % pp cut-off) than the uncorrected Naive Empirical Bayes method,

that is also implemented in `codeml` (Zhang, Nielsen, and Yang, 2005). As BEB depends on the priors from the ML estimation, it is sensitive to inaccuracies in the ML parameters (Anisimova, Bielawski, and Yang, 2002). With highly similar sequences, the predictive power of BEB might be too low to point to individual positions under positive selection (Anisimova, Bielawski, and Yang, 2002).
To apply the codon substitution models described in section 1.4.1 to given codon alignments, the following assumptions should be met (Baker et al., 2016):

- no gene conversion/recombination
- stable ML estimation
- robustness of results when testing different codon (frequency) models
- accurate alignment of homologous codon positions

High rates of gene conversion/recombination within genes of interest might introduce a high rate of false positives (Anisimova, Nielsen, and Yang, 2003). These effects are excluded in the current work by either excluding the columns subjected to gene conversion from the alignment or by excluding the full-length sequence. I do not expect recombination or gene conversion to take place between different deuterostome species, but only between similar paralogs within the same species. Nevertheless, effects detected between species could be caused by incomplete lineage sorting of closely related species such as bonobo and chimpanzee (Manuel et al., 2016). In contrast, high differences in the GC-content has not been found to be a confounding factor in branch-site models even when including genes of cold- and warm-blooded vertebrates (Zhai et al., 2012; Gharib and Robinson-Rechavi, 2013). Recently, Bielawski, Baker, and Mingrone (2016) published a method, that enables assessment of the sampling error during the natural selection test and disclosure of violations of codon model requirements. The `CODEML_SBA` method estimates model parameters on bootstrapped replicates of the original alignment providing "confidence intervals" similar to BS values. Another critical factor for the inference of natural selection is sequence divergence. Sequences with a low divergence do not carry many informative sites, while sequences with very high divergence might be difficult to align and thus violate model assumptions and increase the rate of

Table 1.2: Parameters in branch-site models of natural selection used in the current study as implemented in `codeml` (Yang and Nielsen, 2002; Zhang, Nielsen, and Yang, 2005). Abbreviations: FG – foreground branch; BG – background branch.

Model name	Site class	Parameters FG	Parameters BG	Comments
A H_0	0	$p_0, 0 < \omega_0 < 1$		purifying selection
	1	$p_1, \omega_1 = \omega_2 = 1$		neutral selection
	2a	$0 < \omega_0 < 1$	$\omega_2 = 1$	purifying selection in FG, neutral selection in BG
	2b	$\omega_1 = \omega_2 = 1$		neutral selection in FG and BG
A H_1	0	$p_0, 0 < \omega_0 < 1$		purifying selection
	1	$p_1, \omega_1 = 1$		neutral selection
	2a	$0 < \omega_0 < 1$	$\omega_2 \geq 1$	purifying selection in BG, positive selection in FG
	2b	$\omega_1 = 1$	$\omega_2 \geq 1$	neutral selection in BG, positive selection in FG

false positives (Fletcher and Yang, 2010). Gharib and Robinson-Rechavi (2013) and Bielawski, Baker, and Mingrone (2016) suggest that the dS rate might reach saturation with highly divergent sequences (individual branch length ML estimate > 3), which might lead to a loss of detection power of the branch-site test. Other critical parameters that highly influence the power of the natural selection analysis are the selective pressure (Zhai et al., 2012), tree topology, number of codon sites (length of the alignment) and the number of sequences and taxa sampled (Anisimova, Bielawski, and Yang, 2002; Bielawski and Yang, 2003). The big emphasis on homology detection and consideration of even fragmented gene loci in the current work greatly contributes to the completeness of the individual sequences. Alignment gaps are non-informative for the analysis in `codeml`. This lead to the exclusion of either the codon position from all homologous sequences in the alignment resulting in the shortening of the alignment, or the exclusion of the full-length sequence resulting in a decrease of homologs included. Both factors are known to decrease the power of the natural selection analysis and thus give more power to analysis of curated annotations of protein-coding genes (Zhai et al., 2012).

1.4.3 Detection of specificity determining positions

While fully conserved positions within a protein family define functional key features that are shared across the protein family e. g. a common fold or activation mechanism, positions that systematically differ across groups (e. g. paralogs) might be related to functional specificity (Rausell et al., 2010). Those SDPs are important to understand the diversification of a biological function (Chakraborty and Chakrabarti, 2015). Subfamily specific residues might mediate the group's functionality such as ligand binding, protein binding or allosteric regulation (Juan, Pazos, and Valencia, 2013). During recent years, a mature arsenal of different tools appeared that use very different approaches to identify SDPs from a MSA e. g. entropy, amino acid similarity/ physico-chemical properties, 3D structure, machine learning techniques (please see Chakraborty and Chakrabarti (2015) and Chagoyen, García-Martín, and Pazos (2016) for reviews). All discussed methods have been applied to different protein families of interest e. g. Ras GTPase (Pazos, Rausell, and Valencia, 2006; Ye et al., 2008) or the Smad transcription factors (Ye et al., 2008; Brandt, Feenstra, and Heringa, 2010) to detect ligand and protein binding specificities and helped in designing protein mutants that were tested experimentally. Moreover, the SDP tools `S3det` and `Xdet` have been used to characterize residues involved in oligomer formation and catalytic binding activity, respectively. Most SDP methods easily identify SDPs, that are consistently conserved and differ among subfamilies (residue type I), while positions that are variable in one (residue type II) or all subfamilies (marginal conserved residues, MC) are frequently underrepresented (Chakraborty and Chakrabarti, 2015). Ensembl approaches that combine several SDP prediction methods retrieve more reliable predictions (Brandt, Feenstra, and Heringa, 2010; Chakraborty and Chakrabarti, 2015; Chagoyen, García-Martín, and Pazos, 2016). Following along this line in this dissertation, I overlap sets of SDPs predicted by four different methods explained in the following (Chapter 3). SDP detection methods that rely on phylogenetic trees have not been applied in this work to avoid false positives caused by convergent evolution. The `Sequence Harmony` (SH) method calculates the relative entropy between groups (such as paralogs) i. e. focuses on the amino acid composition differences among groups. Amino acid replacement matrices are not considered. The SH score is based on Shannon's general entropy and was adapted to calculate the relative entropy of one group (A) in relation to all groups (Z). The SH score for one group (A) in

comparison to the other groups at one position i of the MSA is calculated as follows (Eq. 1.16, Pirovano, Feenstra, and Heringa (2006) and Brandt, Feenstra, and Heringa (2010)).

$$SH_i^A = \sum_k p_{i,k} \log_b(\frac{p_{i,k}^A}{\sum_{B \in Z} p_{i,k}^B}) \tag{1.16}$$

with b being the minimal amino acid alphabet size and $p_{i,k}$ being the frequency of amino acid type k at position i, respectively. The group SH scores, SH_i^A for all Z, are averaged over the total number of groups N (Eq. 1.17).

$$SH_i = \frac{1}{N} \sum_{A \in Z} SH^A \tag{1.17}$$

The SH score has a range of [0,1] with 0 indicating no shared amino acids across groups and values close to 1 indicating many shared amino acids across groups. MSA columns with a low SH score are potentially SDPs (Pirovano, Feenstra, and Heringa, 2006). Z-scores of 100 random permutations of group labels are calculated by the SH-webserver and can be used as an additional filter to tune the program's performance (Brandt, Feenstra, and Heringa, 2010).
Similarly to `SH`, `Xdet` compares the distribution of the residue composition between groups. Specifically, `Xdet` compares the mutational behavior (patterns of amino acid changes) under consideration of an amino acid replacement matrix at every position of the MSA with *a priori* functional information (group division). The amino acid replacement matrix (with similarity values A) and the functional similarity matrix (with similarity values F) are compared by calculating a Spearman rank-order correlation coefficient (Pazos, Rausell, and Valencia, 2006) for sequences r and s at position i (Eq. 1.18).

$$r_i = \frac{\mathrm{cov}(A_{rsi}, F_{rs})}{\sigma(A_{rsi})\sigma(F_{rs})} \tag{1.18}$$

A high correlation coefficient r indicates that this MSA column/position characterizes the functionality well. Unsupervised `Xdet` assumes that the functional classification is represented by the overall sequence similarity. This method can identify functional positions for which the classification is not implicit on the alignment or phylogenetic tree (Pazos, Rausell, and Valencia, 2006).
As `Xdet`, `S3det` can be run unsupervised to simultaneously define groups (subfamilies) and to identify the corresponding SDPs (Rausell et al., 2010). `S3det` is based on multiple correspondence analysis (MCA), a technique for analysis of multivariate data that is similar to principal component analysis, but applied to categorical data. `S3det` first represents the input MSA as a binary matrix encoding the amino acid identity at a specific position. The coordinate system that displays these initial vectors is then transformed so that the principal axes (eigenvalues) represent the sources of variation of sequences vs. residue-positions. The MCA accomplishes an orthogonal decomposition of those sources of variation. Next, groups of sequences are identified by a k-means clustering approach. Residue positions are assigned to the nearest sequence groups. SDPs are identified by ranking the residue positions by distance to the principal axes. For more details about the method, please refer to the Supplemental Information of Rausell et al. (2010).
The feature-weighting machine learning algorithm `multi-RELIEF` (Ye et al., 2008; Brandt, Feenstra, and Heringa, 2010) does not take into account amino acid similarity either. While iterating over all sequences l from two groups and the positions i within

the MSA, it updates a weighting-vector of alignment length (initialized as 0) based on the position's ability to distinguish the nearest neighbor from a different group and the same group (Eq. 1.19). The nearest neighbor of one sequence r in the same or different group, respectively, sequences `miss(r)` and `hit(r)`, are defined as the sequences with the minimal number of mismatches between r and any sequence of the respective group. Position-specific weights (w_i) between a pair of groups are updated as follows with d being the Hamming distance:

$$w_i = w_i + d(r, miss(r)) - d(r, hit(r)) \tag{1.19}$$

The position-specific weights are averaged across the number of sequences in the current group. In the multi-group implementation (Ye et al., 2008; Brandt, Feenstra, and Heringa, 2010; Brandt, Feenstra, and Heringa, 2016), the pairwise positive and negative position-specific weights are calculated for all group pairs and averaged across the number of positive and negative weights, respectively. The resulting `multi-RELIEF` values have values between [-1,1]. Residues that are conserved in all groups, but discriminate between groups have a positive value (residues of type I), while residues that are divergent within groups and conserved across groups have negative weights (Brandt, Feenstra, and Heringa, 2016).

1.5 Multi-talents in cell signaling: The cytosolic arrestin proteins

Arrestins are a very interesting protein family as they have many different interaction partners due to their function as early signaling relay and scaffolding proteins illustrated in detail below. The detailed exploration of arrestin evolution in deuterostomes is one of the main results of this thesis presented in Chapter 3.

1.5.1 Functions of arrestins in cell signaling

Communication and reaction to extra-cellular stimuli are prerequisites for the survival of every living cell. An important class of proteins that can receive and transduce extra-cellular signals such as small molecules, peptides, nucleotides, odorants or photons, are the seven trans-membrane GPCRs (Bockaert and Pin, 1999). GPCRs undergo a conformational change upon extra-cellular ligand (agonist) binding leading to the recruitment of the heterotrimeric G protein complex towards its C-terminus at the intracellular, cytosolic side (Fig. 1.8). The active GPCR triggers Gα protein activation by opening of the guanine nucleotide binding pocket allowing for an exchange of guanosine diphosphate (GDP) with guanosine triphosphate (GTP). This conformational change results in the dissociation of the Gα and Gβ/γ subunits from the activated GPCR (Alberts, 2011). This process can initiate different downstream signaling pathways that control key cellular processes such as apoptosis, proliferation or differentiation, mainly by repression and activation of transcription. Phosphorylation of the activated GPCR C-terminus by a G protein receptor kinase initializes a feed-back loop of GPCR desensitization (Krupnick and Benovic (1998), section 1.1.4). Another key-player of the fast and precise shut-off of GPCR signaling via G proteins is the cytosolic arrestin protein with a molecular weight of about 40-45 kDa (Lohse et al., 1990). Arrestins preferentially bind to activated and phosphorylated GPCRs by blocking their inter-helical cavity, thereby precluding its coupling to cognate G proteins (Fig. 1.8, Gurevich and Gurevich (2006b) and Kang et al. (2015)). In particular, arrestin binding is indispensable for a high temporal resolution in vision (Renninger,

Gesemann, and Neuhauss, 2011; Gurevich et al., 2011). Vision is mediated by a subgroup of GPCRs, the visual opsin receptors, which covalently bind a chromophore. The chromophore undergoes a light-induced isomerization precluding a subsequent activation by another photon before chromophore regeneration. Diverse other biological functions of arrestins have been described in the last two decades, that go beyond their "arresting"–function that gave the protein family its name. Among them are scaffolding, subcellular localization, and regulation of kinases, phosphodiesterases and ubiquitin ligases, cytoskeletal reorganization, G protein independent signaling and GPCR trafficking (for reviews please see Luttrell, 2013; Gurevich et al., 2014).

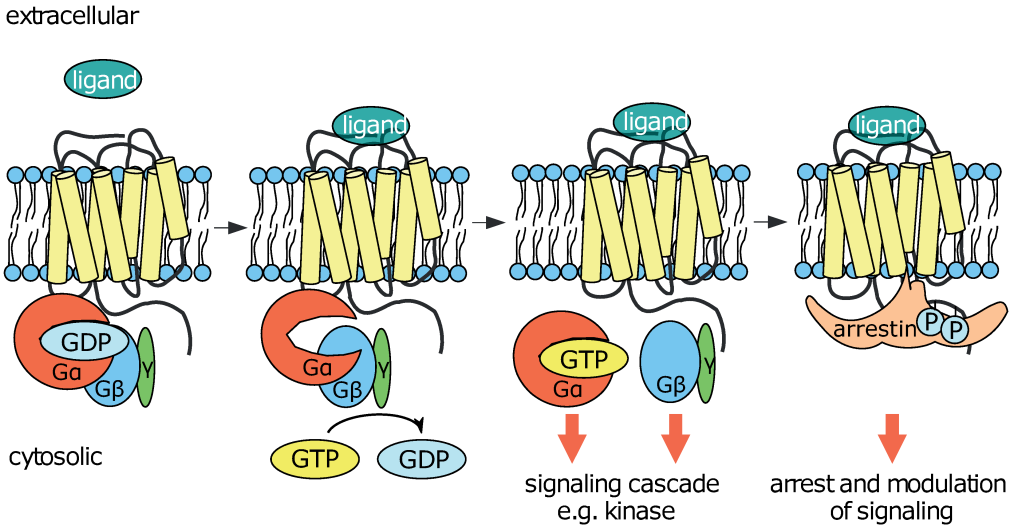

Figure 1.8: Role of G proteins and arrestins in GPCR signaling. Schematic depiction of the transmembrane G protein-coupled receptor (GPCR, yellow) situated in the cellular membrane of an eukaryotic cell together with its cytosolic binding partners, G proteins and arrestin. The GPCR undergoes a conformational change upon ligand-induced activation and recruits the heterotrimeric G protein (Gα, β, γ: orange, blue, green). The Gα protein alternates between an active (GTP-bound) and inactive (GDP-bound) state. Downstream signaling is initiated by GTP-hydrolysis, which mediates activation of effectors that can catalyze release of second messengers such as cyclic adenosine monophosphate, inositol-1,4,5-triphosphate or calcium. Integration of those signals regulates final effector proteins (transcription activators and repressors). One downstream effector is the G protein receptor kinase, which phosphorylates the GPCR C-terminus and thus increases the binding affinity of arrestin (peach), which competes with G proteins for binding of the active GPCR. Please note that this depiction is simplified and conformational changes are not shown.

There are four arrestin paralogs in mammals, functionally divided into the visual and non-visual group, each composed of two members. The visual arrestin-1 (formerly known as rod arrestin) is well known for binding to phosphorylated rhodopsin with high specificity and affinity, preferring it over other GPCRs (Vishnivetskiy et al., 2004; Vishnivetskiy et al., 2011). The same paralog represents the most prevalent arrestin expressed in mouse cones, suggesting that it binds to cone pigments (Nikonov et al., 2008). In contrast, binding specificity of the second visual paralog, arrestin-4 (formerly known as cone arrestin or X-arrestin), is ensured by its co-expression with cone opsins in cone photoreceptors, as *in vitro* arrestin-4 binds non-visual GPCRs fairly well (Sutton et al., 2005). Both visual arrestins were hypothesized to be indispensable for high resolution color vision under bright light due to their complementary properties in regard to oligomerization and receptor-binding affinity (Gurevich et al.,

2011). In contrast, the non-visual arrestins, arrestin-2 and arrestin-3 (also known as β-arrestin1 and β-arrestin2), have a broad receptor specificity recognizing several hundred different GPCRs and are ubiquitously expressed. While there are overlaps in cell-type, cell-compartment expression (Neuhaus et al., 2006; Hoeppner, Cheng, and Ye, 2012) and binding capability to several interaction partners, e. g. the MAP kinase kinase MKK4 (Zhan et al., 2011b) or the class B GPCRs (Oakley et al., 2000), differences between the non-visual arrestins exist in regard to, among others, concentration (Gurevich, Benovic, and Gurevich, 2002), receptor selectivity profiles (Oakley et al., 2000) and other interaction partners (Xiao et al., 2007). Arrestin-3 is the least selective member of the arrestin family with lower preference for active phosphorylated receptors over the inactive form (Gurevich et al., 1995; Zhan et al., 2011a), while it displays higher affinity for class A GPCRs than arrestin-2 (Oakley et al., 2000). Another example of a paralog-specific downstream effect is activation of JNK3 promoted by arrestin-3 specifically (Song et al., 2009). Binding of phosphoinositides, e. g. inositol-hexa-sphosphate (IP6) facilitates oligomerization of non-visual arrestins, while it inhibits oligomerization of arrestin-1 (Milano et al., 2006; Hanson et al., 2008). Given the importance of GPCRs in diseases like cancer (Lappano and Maggiolini, 2017), multiple sclerosis (Du and Xie, 2012), Alzheimer's disease (Thathiah and Strooper, 2011) and obesity (Kimple et al., 2014), among others, it is not surprising that they are targets of 25-30 % of all drugs of pharmaceutical industry today (CHI, 2017). Biased signaling, the agonist-induced, selective stabilization of a specific GPCR conformation, has the potential to favor a specific arrestin or G protein conformation thereby mediating specific downstream-signaling pathways. In recent years, considerable efforts were made towards the design of arrestins that modulate GPCR signaling and facilitate biased signaling (Liu et al., 2015; Cahill et al., 2017; Zhou, Melcher, and Xu, 2017).

1.5.2 Arrestin activation by G protein-coupled receptor binding

Arrestin proteins are composed of two domains each with the β-sandwich at its core, the *arrestin_N* and *arrestin_C* domain (section 1.1.4, Fig. 1.9 A, B). The N-domain contains the only α-helix. A highly flexible linker region connects both domains at the central crest (Fig. 1.9 C, Zhan et al. (2011a)). Representatives of all four orthology groups have been crystallized (Hirsch et al., 1999; Han et al., 2001; Sutton et al., 2005; Zhan et al., 2011a) and reveal an overall similar fold and activation mechanism, despite the in-detail functional differences explained above. Arrestin-1 binding to activated and phosphorylated rhodopsin is the model system used to study arrestin activation through crystallization, mutagenesis and functional assays (Tab. 1.3). The basal, inactive state of arrestins is characterized by an intact polar core and a hydrophobic interaction between β-strand I of the N-terminus, α-helix I and β-strand XX of the arrestin C-tail (three element interaction, Fig. 1.9 B). The polar core interaction is untypical for a soluble protein as it buries six charged residues within the protein core (section 1.1.4, Hirsch et al. (1999)). GPCRs engage the concave side of arrestins (Hanson et al., 2006; Hanson and Gurevich, 2006; Vishnivetskiy et al., 2011) leading to the replacement of the arrestin C-tail by the phosphorylated C-terminus of the receptor (Kang et al., 2015) resulting in the disruption of the three element interaction and the release of the arrestin C-tail. Positively charged lysine and arginine residues at the arrestin N-terminus bind to the receptor phosphates first (Granzin et al., 2015; Kang et al., 2015) and deliver them to the polar core residue R175 that gets accessible upon movement of the lariat loop (D296-N305) in the central crest region (Fig. 1.9

C, inlet). This results in the disruption of the salt bridge R175-D296 and the destabilization of the polar core (Han et al., 2001) followed by a 20 ° rotation movement of both domains relative to one another (Kang et al., 2015). Charge reversal of one of the residues engaged in the salt bridge (e. g. R175E, Granzin et al. (2015)), a triple Ala mutation of the three element interaction (Kang et al., 2015) and the ablation of the arrestin C-tail as naturally occurring in the p44 isoform of arrestin-1 (Kim et al., 2013) thus all result in the "pre-activation" of arrestin as described above and render those arrestin mutants insensitive to the phosphorylation state of the receptor. The receptor C-tail forms an extended β-sheet with strand IV of arrestin accompanied by the re-orientation of the middle and lariat loops and an elongation of the finger loop in the central crest (Shukla et al. (2013) and Kim et al. (2013), Fig. 1.9 C, inlet). Those regions directly interact with the active receptor precluding G protein binding (Kim et al., 2013; Szczepek et al., 2014; Kang et al., 2015). Kim et al. (2013) further hypothesized that the inter-domain rotation facilitates the adaptive fit of arrestin to the active receptor. Another region that strongly reduced receptor binding upon mutation is the C-edge that was speculated to interact with membrane phospholipids by Ostermaier et al. (2014) and Kang et al. (2015). Although described as a two-step process here (release of the arrestin C-tail and recognition of the receptor active state) according to Ostermaier et al. (2014), recent studies open up the possibility that some receptors (e. g. the vasopression type 2 receptor) are stably bound by the arrestin C-tail omitting the finger loop interaction (Cahill et al., 2017).

Table 1.3: Key functional elements in arrestin activation and receptor binding mapped in reference to cow arrestin-1.

Function	Residues	Reference
Polar core	D30, D33, R175, K176, D296, D303, R382	Hirsch et al. (1999)
Three element interaction		Han et al. (2001) and Luttrell (2013)
β-strand I (N-terminus)	V11, I12, F13	
α-helix I	L103, L107, L111	
β-strand XX (C-terminus)	F375, V376, F377	
Receptor binding		Ostermaier et al. (2014)
phosphate sensor	K14, K15, R18, K20, R29, K110, K166, K300	
finger loop	Q69, D73–M75	
lariat loop	L249–S252, Y254	
C-edge	W194–S199, K232, G337–G340, T343-S345	
middle loop/loop 139	Q133-S142	

1.5.3 Functions of arrestins in cellular trafficking

Non-visual arrestins mediate internalization of GPCRs and are directed to the cellular membrane upon agonist binding, first shown for the β-2 adrenergic receptor (Goodman et al. (1996), Fig. 1.10). Non-visual arrestins are involved in the mechanism that determines whether receptors are recycled or degraded after endocytosis (Shenoy and Lefkowitz, 2003). The elimination of the receptor from the membrane furthermore regulates the sensitivity of cellular response towards ligand binding. Upon activation,

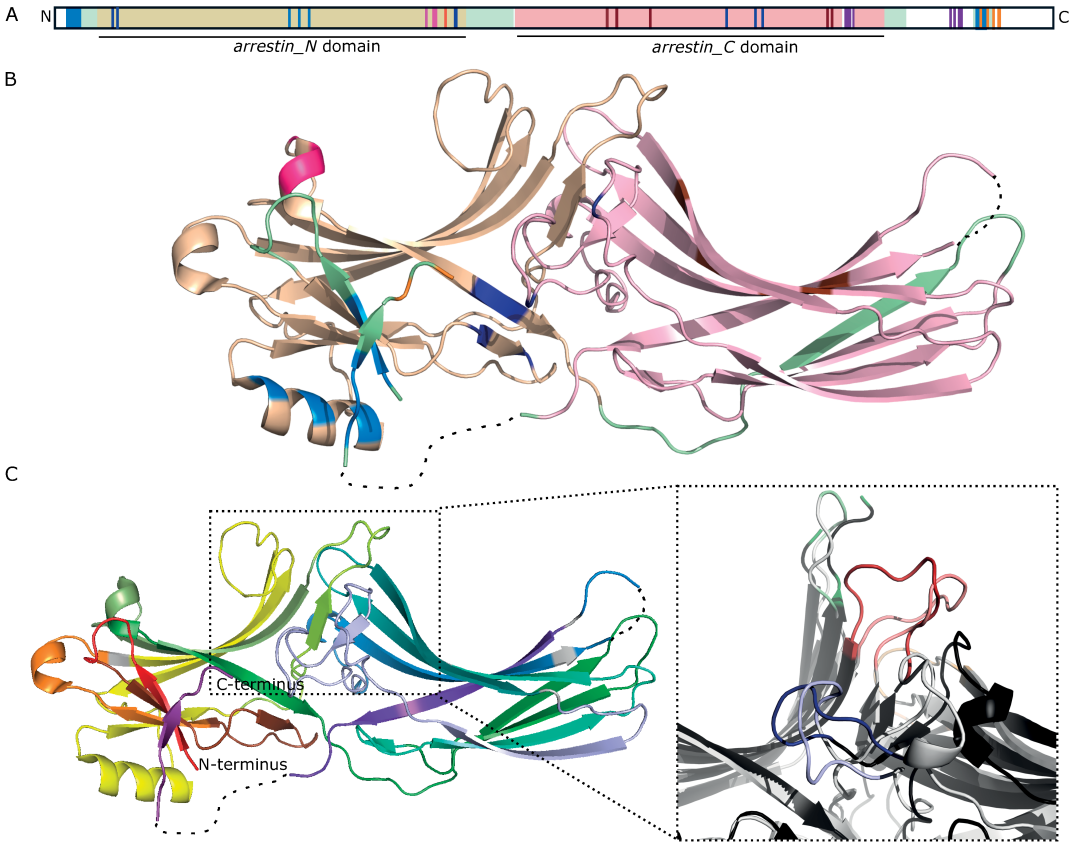

Figure 1.9: Functional elements of arrestins. A – Schematic, linear representation of bovine *ARRB1* with functional elements in bright colors (orange – AP-2 binding site, light blue – three element interaction, dark blue – polar core, brown – high affinity IP6 binding site, pink – low affinity IP6 binding site, red – phosphate sensor, purple – clathrin binding sites). Arrestins encode two key domains, the *arrestin_N* domain (wheat) and the *arrestin_C* domain (light pink). Other regions are shown in light green, while sequence stretches that are not visible in the crystal structure (B) are shown in white. B – Crystal structure of bovine *ARRB1* in the inactive state with functional features colored according to A (PDB: 1G4R (Han et al., 2001)). The clathrin binding sites are missing in the crystal structure (dotted lines). C – Same as B colored according to the conserved exon-borders in vertebrates (rainbow coloring from exon 2 – red to exon 16 – dark violet). Exons 1, 15 as well as parts of exons 13, 14 and 16 are missing in the crystal structure (shown as dotted lines if not situated on the N- or C-terminus). Amino acids whose codons are split among two exons are shown in grey (Flicek et al., 2014). Inlet: Overlay of the central crest region of bovine full-length (inactive) *SAG* (light color, PDB: 1CF1 (Hirsch et al., 1999)) and bovine p44 (active) *SAG* (dark color, PDB: 1G4R (Kim et al., 2013)). Activation induces movements of the finger loop (green), loop 139 (red), the lariat loop (blue, front), the gate loop (orange, back) and a domain rotation. Crystal structure images were created with `Pymol 1.8.4.0 Open-Source` (Schrödinger, 2015).

non-visual arrestins interact with the two most abundant protein components of clathrin-coated vesicles, clathrin and the heterotetrameric adapter complex AP-2 (Kirchhausen, Owen, and Harrison (2014), Fig. 1.9 B, C).
The non-visual vertebrate arrestins and visual arrestin-1 from fly harbor a homologous binding site in the arrestin C-tail that follows the strict consensus motif "[E/D]xxFxx[F/L]xxxR" (Laporte et al., 2000; Milano et al., 2002; Schmid et al., 2006).

The respective motif adopts an α-helical conformation that contacts the β-appendage of AP-2 on its top side as seen in the crystal structure of the β-AP-2 appendage and the respective arrestin peptide (Schmid et al., 2006; Moaven et al., 2013). The α-helix is not formed until arrestin activation with the residues "IVF" situated immediately neighboring to the AP-2 consensus motif in arrestin negatively regulating the interaction (Burtey et al., 2007). Binding of vertebrate arrestin-1 to the β-appendage is much weaker compared to arrestin-2 due to a mutation in the respective consensus motif (Schmid et al., 2006; Moaven et al., 2013). Nevertheless, this interaction might be relevant due to arrestin-1's high concentration in rods (Moaven et al., 2013). Non-visual arrestins interact with another component of the heterotetrameric AP-2 protein complex, μ adaptin. This interaction is mediated by the short motif "[Y/F]VTL" situated on the concave site of the *arrestin_N* domain and regulated by phosphorylation of the tyrosine residue (Marion et al., 2007).
Non-visual arrestins bind the other main component of the endocytosis machinery, clathrin, by interaction with clathrin's N-terminal β-propeller domain. *ARRB1*'s C-terminus contains two clathrin binding sites (CBS). They are referred to as the major and minor CBS due to their different binding affinities to clathrin (Kang et al., 2009). The major CBS interacts with blades 1 and 2 of clathrin as do other proteins of the endocytosis machinery like the β-subunit of AP3, 2 and 1 (ter Haar, Harrison, and Kirchhausen, 2000), while the minor site interacts with the shallower groove between blades 4 and 5 (Kang et al., 2009). This mainly hydrophobic interaction does not require any specific orientation and thus might allow flexibility in the macromolecular assembly of different components of the endocytosis machinery. The minor CBS is located on exon 13, which does not exist in *ARRB2* and can be excluded by exon skipping in *ARRB1*. One CBS is sufficient for receptor internalization given an intact AP-2 motif (Burtey et al., 2007; Kang et al., 2009). Kang et al. (2009) showed in a pull-down assay with immobilized clathrin that the full-length *ARRB1* with both CBSs binds about 50 % more clathrin than the shorter isoform lacking the minor CBS. As with AP-2 binding, the basal conformation of arrestin does not bind clathrin (Kern, Kang, and Benovic, 2009).
The stochiometry and macromolecular arrangement of clathrin, different arrestin isoforms, arrestin paralogs and AP-2 is not well understood even when ignoring endocytosis accessory proteins that compete with arrestins for the same binding sites on AP-2 and clathrin (Laporte et al., 2002; Kang et al., 2009). Post-translational modifications and additional binding partners influence arrestin-mediated endocytosis, e. g. N-ethylmaleimide-sensitive factor, ARF6, PI4P kinase or phosphoinosites (Shenoy and Lefkowitz, 2011). Binding of the phosphoinosite IP6 to arrestin-1's low affinity binding site triggers the release of the C-terminus during a conformational change towards the active state (Zhuang et al., 2010). Both non-visual arrestins possess two IP6 binding sites, a high and a low affinity IP6 site. Nevertheless, IP6 can have opposing effects on the interaction of *ARRB1* and *ARRB2* with clathrin and the receptor depending on its concentration (Gaidarov et al., 1999). Arrestin-3 mutants with a disrupted high affinity IP6 binding site fail to mediate internalization of the β-2 adrenergic receptor via endocytosis (Gaidarov et al., 1999; Tian, Kang, and Benovic, 2014) adding another level of complexity to endocytosis regulation with involvement of arrestins.

1.5.4 Evolution of arrestins

Arrestin proteins belong to the arrestin clan and were named β-arrestins by Alvarez (2008) or true arrestins by Gurevich and Gurevich (2006a), Aubry and Klein (2013),

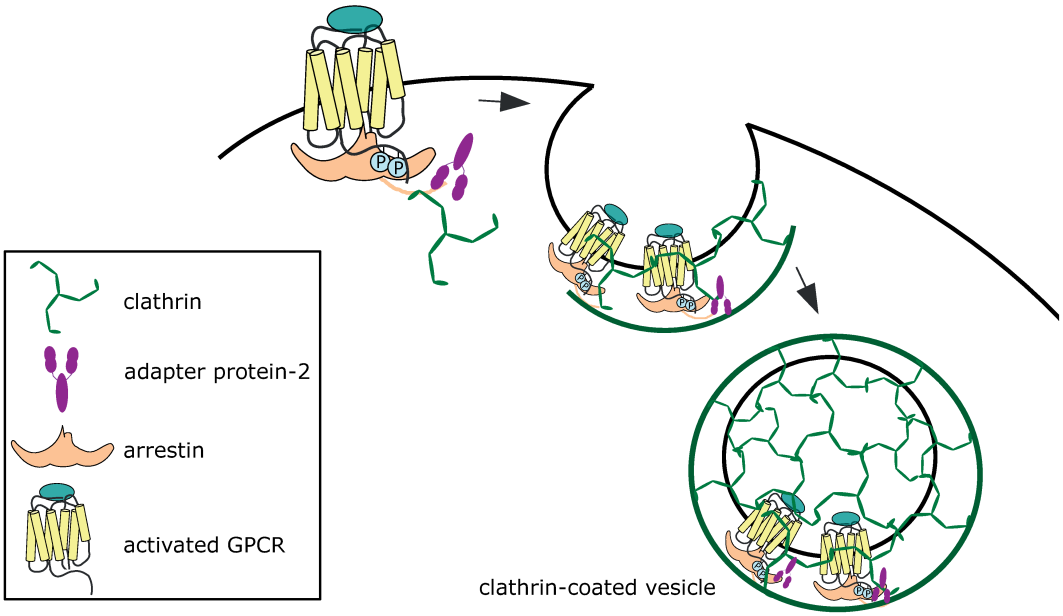

Figure 1.10: Non-visual arrestins mediate endocytosis. Activation of arrestins results in the release of its C-terminus, which contains two clathrin binding and an adapter protein-2 binding site. Binding of clathrin and adapter protein-2 initialize the formation of clathrin coated vesicles. The receptor might be degraded or recycled depending on subsequent trafficking of the vesicles. Other accessory proteins that are also involved in coated vesicle formation and vesicle budding are not shown.

and Kang et al. (2015). Below, I will refer to this group of proteins as arrestins although there are additional members in the clan that share the two domain composition of Ig-like β-strand sandwich folds with possible insertions and extensions (Shi et al., 2006; Collins et al., 2008; Aubry and Klein, 2013; Polekhina et al., 2013; Sonoda, Mizutani, and Mikami, 2015). The arrestin clan members do not share a common function, although all are connected to trafficking and scaffolding (Aubry and Klein, 2013). These are the arrestin-domain containing proteins with 11-17 % sequence identity to arrestins (Aubry, Guetta, and Klein, 2009) and a set of families that are even more distantly related to arrestins with maximal 10 % sequence identity (Aubry and Klein, 2013). These distant relatives encompass the Vacuolar protein sorting protein 26 (Vps26) family, DSCR3, and RGP1 that are represented in human, as well as fungal arrestin-related trafficking adapters, amoebal arrestin domain-containing proteins and the Spo0M family in Bacteria and Archaea (Alvarez, 2008; Aubry and Klein, 2013). The shared fold of the arrestin clan was hypothesized to be the result of convergent evolution (Aubry and Klein, 2013). In contrast to the rest of the arrestin clan, the sequences of arrestins are highly conserved (Luttrell, 2013).
The human genome entails four arrestin paralogs. The genes *SAG* and *ARR3* encode the visual arrestins, arrestin-1 and arrestin-4, while the genes *ARRB1* and *ARRB2* encode the non-visual arrestins, arrestin-2 and arrestin-3, respectively. Both functional groups seem to be monophyletic (Gurevich and Gurevich, 2006a). Visual arrestins exhibit a much higher evolutionary rate than non-visual arrestins (Hisatomi et al., 1997; Kawano-Yamashita et al., 2011). Arrestins have been found in all holozoan clades except for Ichthyosporea: Choanoflagellates, Filasterea and animals (Mendoza, Sebé-Pedrós, and Ruiz-Trillo, 2014). Within bilaterians, the clade of animals with

most living representatives, arrestins are found in both deuterostomes and protostomes (Gurevich and Gurevich, 2006a; Alvarez, 2008; Aubry, Guetta, and Klein, 2009; Mendoza, Sebé-Pedrós, and Ruiz-Trillo, 2014). Arrestins were studied extensively in mammals in the past (Granzin et al., 1998; Smith et al., 2000; Maeda et al., 2000), although individual arrestins from non-mammalian vertebrates have been cloned for functional studies. Among them are visual arrestins from frogs (Craft and Whitmore, 1995; Abdulaeva, Hargrave, and Smith, 1995; Mani, Besharse, and Knox, 1999), salamander (Smith et al., 2000) and gecko (Zhang, Wensel, and Yuan, 2006). Phylogenetic analyses support a 1:1 orthology with their human counterparts. Co-expression of two distinct arrestin-1 genes, termed *SAGa* and *SAGb*, in rods of medaka was reported by Imanishi, Hisatomi, and Tokunaga (1999). Renninger, Gesemann, and Neuhauss (2011) identified two zebrafish paralogs for each visual arrestin ortholog in human, as well as two zebrafish paralogs for arrestin-3. They concluded that these three additional arrestin genes originated from the teleost-specific 3R-WGD event (section 1.2.1). The arctic lamprey expresses a visual and a non-visual arrestin in its pineal organ (Kawano-Yamashita et al., 2011). Nakagawa et al. (2002) showed that the vase tunicate, has only a single arrestin with functional features of both visual and non-visual subtypes. This suggests that the divergence of visual and non-visual arrestins is indeed associated with the vertebrate-specific 2R-WGD (section 1.2.1). A comprehensive phylogenetic analysis to test this hypothesis, however, still has been missing.

Chapter 2

The `ExonMatchSolver`-pipeline – gene annotation on fragmented assemblies

This Chapter is based on Indrischek et al. (2016).

2.1 Motivation

Accurate multiple sequence alignments (MSAs) are required as input for a wide variety of different computational analysis techniques, e. g. in phylogenetics, molecular evolution and comparative genomics (section 1.4). Tests for inter-residue co-evolution (Juan, Pazos, and Valencia, 2013) and correlation of conservation with protein structure (Celniker et al., 2013) allow for identification of functional motifs and elements. Protein interfaces and interaction partners can be predicted considering inter-protein co-evolution (Juan, Pazos, and Valencia, 2013). These approaches can be used to improve protein structure prediction. Sequence alignments also form the basis for evaluating changes in natural selection pressures over evolutionary time scales (Nowick et al., 2011).
Many large protein families, such as transcription factors, growth factors, proteins involved in signaling pathways or membrane proteins include paralogous members that share highly similar sequence elements. Detailed phylogenies of these protein families – usually referred to as gene trees – are utilized to reveal rapid gene loss and pseudogenization, frequent gene duplication and abundant gene conversion events (Cortesi et al. (2015), section 1.4.1). The reconstruction of accurate gene trees for protein families, however, has turned out to be one of the most recalcitrant problems in computational biology. This has multiple causes. One key issue, which is the main motivation for the work of this Chapter, is the availability and quality of the input sequence data (section 1.3). Sequences extracted from databases are usually incomplete and may contain annotation errors.
I address this particular issue here by describing an algorithm that identifies the optimal assignment of coding exons to genomic fragments. In contrast to existing methods, which find, separately for each query paralog, the best match(es) in the genome, the developed tool, the `ExonMatchSolver`-pipeline (EMS-pipeline) identifies the collectively best match of an entire group of highly similar paralogous genes to a set of genomic loci.

2.2 Methods

2.2.1 Pipeline overview

In the following, I will refer to the protein subsequence encoded by the coding portion of exon x as TCE x (Translated Coding Exon), i. e. the conceptual translation of the protein-coding subsequence of the corresponding exon. In this work, it is assumed that the entire group of paralogs admits a hypothetical "ancestor" from which each of the proteins can be derived by deletion of TCEs. TCE x is thus homologous to all TCEs x from other paralogs. The number of paralogs to be identified in the target genome is either assumed to be identical to the number of family members encoded in the query genome and provided as input file or can be specified by the user. The EMS-pipeline implements a work-flow comprising four main steps: (1) The search of protein sequences or protein-models specific for paralogs and individual TCEs against a complete target genome; (2) The paralog-to-contig assignment formulated as an Integer Linear Programming (ILP) problem, (3) a refined search for exons missing after step 2 relative to the input gene models, and (4) the assembly of fragmented hits and the proposition of gene annotations. The formulation of the ILP is the core of the EMS-pipeline and will be referred to as `ExonMatchSolver` in the following. The EMS-pipeline produces both a predicted protein sequence for each paralog, and an assignment of each predicted paralog to a paralogous group. The EMS-pipeline accommodates several types of input. If paralog-specific and individual-TCE alignment-files are provided, profile Hidden Markov Models (pHMMs) are built (0a) and used as queries. Otherwise, homologous TCE groups across paralogs within the query genome can be identified in an additional pre-processing step (0b). The overall organization of the underlying workflow is summarized in Fig. 2.1.

2.2.2 Exon assembly as an assignment problem

The key difficulty is the creation of a complete and accurate gene model of the coding sequence on fragmented genome assemblies. The starting point is a set $\{Q_1, \ldots, Q_N\}$ of N paralogous query proteins. Each query protein Q_j can be decomposed into its TCEs $(q_1^j, q_2^j, \ldots, q_{m_j}^j)$. For a set $\{X_1, X_2, \ldots, X_n\}$ of contigs, a similarity score θ_{ijk} measures how well TCE q_k^j of paralog j matches to contig i. Fig. 2.2 illustrates the problem setup.

The term contig here is used to refer to a genomic locus harboring at most one gene of interest. If the contigs in the genome assembly are very long, they may have to be subdivided so that each target sequence contains only a single locus of interest e. g. by creation of a new, artificial contig that was not contained in the original assembly. Furthermore, all contigs without significant matches are removed before solving the paralog-to-contig assignment problem.

The assumption that each TCE can be derived from a hypothetical "ancestor" by deletion of TCEs covers all gene families in which the gene structure has not undergone permutations of exons. For instance, if an exon was split in one lineage by insertion of an extra intron (intron gain), this extra intron boundary can be traced back to the "ancestor" and inserted within all its descendants. TCEs then have to be artificially split at this boundary. After this preparatory step (which is left to the user in this implementation), the TCE blocks (in the following simply called TCEs for brevity) are numbered consistently, in the sense that homologous TCEs have the same number and $m_j = m$ becomes independent of the paralog. Missing (deleted) TCEs simply remain unmatched. The quality of a match between query TCE q_k^j to a genomic

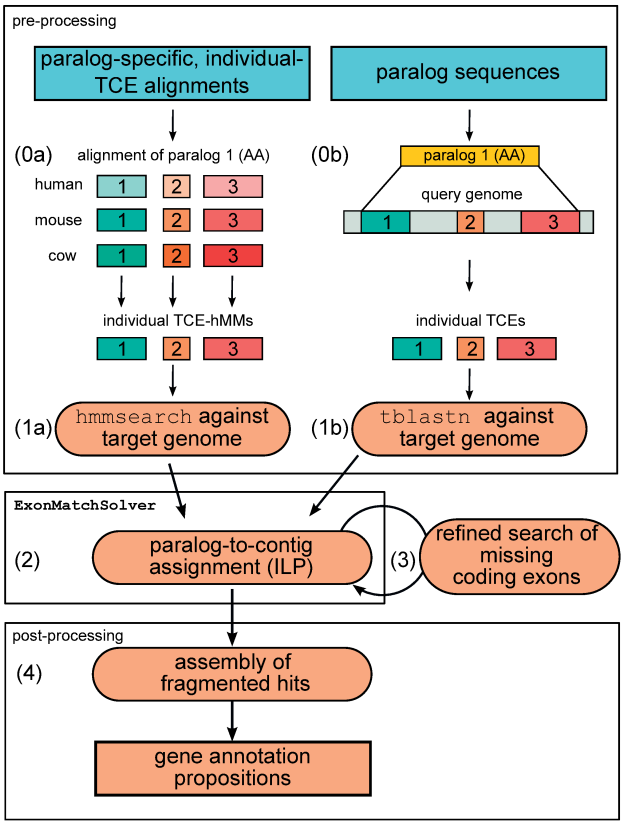

Figure 2.1: The EMS-pipeline explicitly solves the paralog-to-contig assignment problem. Sequence matches to individual TCEs are collected in a step-wise procedure applying either `tblastn` (from single sequences of individual TCEs) or `hmmsearch` (starting from a sequence alignment for each TCE). Depending on the input, pre-processing steps (0a) or (0b) are performed before similarity search. The colored boxes represent TCEs. The pre-processing steps, which are performed separately for all individual TCEs of all paralogs, are exemplified here for one paralog encoded by three exons. For a detailed description of the individual steps, please refer to section 2.2.5. Abbreviations: AA – amino acid sequence; pHMMs – profile hidden Markov Models; ILP – Integer Linear Programming problem; TCE – translated coding exon.

match i in contig X_i is measured by the bit score θ_{ijk} computed by either `tblastn` (Camacho et al., 2009) or `hmmsearch` (Eddy, 2008; Eddy, 2011). To remove spurious hits, an E-value filter is employed first. Secondly, TCE-hits that are found alone on one contig without any accompanying hits are subjected to a length-normalization and a bit score-filtering. Those exons that have a length-normalized bit score above the cut-off (> 1.5) are called "single, reliable exons" in the following. For undesirable assignments, $\theta_{ijk} = 0$.

The paralog-to-contig assignment problem is a combination of a matching problem (Lovász and Plummer, 1986) and an assignment problem (Burkard, Dell'Amico, and Martello, 2012). It can be phrased formally as follows:

Paralog-to-Contig Assignment Problem (PCAP)

> **Instance:** A set Q of n queries ("paralogs"), each of which comprises a non-empty list of TCEs each denoted (j, k) with $1 \leq j \leq n$ and $1 \leq k \leq m_j$; a set T of N targets ("contigs"), each comprising a list of sites (i, h) with $1 \leq i \leq N$ and $1 \leq h \leq M_i$; scores $\sigma_{i,h;j,k}$ measuring the similarity of query TCE (j, k) with target site (i, h).

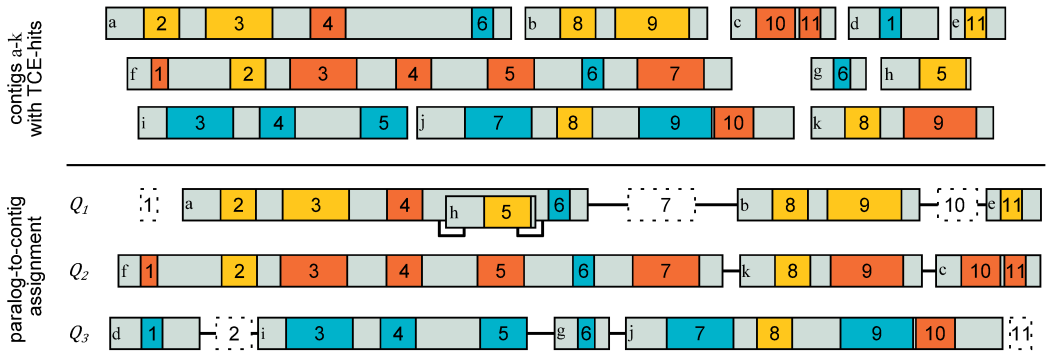

Figure 2.2: Illustration of the paralog-to-contig assignment problem. In this hypothetical example, each of the three paralogous genes has 11 coding exons which are homologous to the respective exon of the other paralogs (numbers 1-11). The paralogs are distributed over 11 contigs of different sizes, which are denoted by letters. TCE-hits on the 11 contigs are colored according to the query paralog Q_i that scored best (yellow – query paralog Q_1, red – query paralog Q_2, blue – query paralog Q_3). The lower part of the figure shows the assignment identified by the EMS-pipeline. Note, that exon 5 of paralog 1 is inserted into contig a, whic carries exons 2-4 and 6. Putative missing (or deleted) exons are shown as dotted boxes. Abbreviation: TCE - translated coding exon.

Solution: A bipartite matching $\mathcal{M}$ of query TCEs (j,k) and target sites (i,h) so that

1. each target i is assigned to at most one query j, i.e., $(j,k):(i,h) \in M$ and $(j',k'):(i,h') \in M$ implies $j' = j$, and
2. if $(j,k):(i,h) \in M$ then there is $(j,k'):(i,h') \in M$ for every TCE k' of the query j for which there is a site h' on the same target i with $\sigma_{i,h';j,k'} > 0$.

The target sites are interpreted as (parts of) exons so that in instances of practical interest to us, each TCE and each site can be assigned a type τ so that $\sigma_{i,h;j,k} > 0$ if and only if $\tau(i,h) = \tau(j,k)$.

Objective Function:

$$f(\mathcal{M}) = \sum_{(j,k):(i,h)\in\mathcal{M}} \sigma_{i,h;j,k} \to \max! \tag{2.1}$$

Assuming for practical applications to biological data, that each exon type appears at most once on each target i, the index h can be suppressed and set $\theta_{ijk} := \sigma_{i,h;j,k}$ if there is (i,h) with $\tau(i,h) = \tau(j,k)$ and $\theta_{ijk} := -\infty$ otherwise. PCAP is a difficult combinatorial optimization problem as shown in the following (proof by Dr. Nicolas Wiesecke and Prof. Peter F. Stadler):

Theorem 1. *The decision problem version of* PCAP *is NP-complete.*

Proof. The PCAP can be reduced from the ***graph 3-coloring*** problem as will be shown below, which is known to be NP-complete (Karp, 1972). Hence, PCAP is NP-complete. Consider an arbitrary graph with vertices (V) and edges (E), $G = (V,E)$, and an associated PCAP with $n = 3$ queries and $m = |E|$ TCEs on each query. For each $i \in V$ one target is created, with $M_i = |\{i' : [i,i'] \in E\}|$ sites. A "type" $\tau \in \mathbb{N}$ is assigned to each query TCE and target site, and set $\sigma_{i,h;j,k} = 1$ if and only if $\tau(i,h) = \tau(j,k)$ and $\sigma_{i,h;j,k} = -\infty$ otherwise, i.e., query TCEs can only match with target sites of the same type. There are $|E|$ distinct types, each associated with a single edge in G. A

target i contains a site of type τ, if and only if the respective vertex is incident to the corresponding edge. Two targets i and i' therefore share a site of the same type if and only if $[i, i'] \in E$. The three queries are constructed as identical lists, each containing TCEs of all $|E|$ types. Therefore, any independent set of targets matches to each query, while no query can match two adjacent targets. A solution of the PCAP constructed in this manner, in which every target is assigned to one of the three queries, implies a 3-coloring of G. Conversely, if a 3-coloring of G exists, it provides a solution of the PCAP.
Finally, it is easy to verify that the PCAP constructed from G has polynomial size: There are $|V|$ targets, each of which has not more than $|E|$ edges, i.e., there are not more than $|V|\,|E|$ target sites and exactly $3|E|$ query TCEs, i.e., the size of the underlying matching problem lives on a graph with $O(|V|^3)$ vertices.
Thus PCAP cannot be easier than graph 3-coloring, which is NP-complete. □

Since Theorem 1 precludes the existence of an efficient solution (unless P=NP), the PCAP is solved by means of ILP. To this end, the formal specification of PCAP from above has to be converted into a set of linear constraints. The notation of similarity scoring is simplified in terms of θ_{ijk}.

2.2.3 Solving the Paralog-to-Contig Assignment Problem

To formulate the PCAP as an ILP, the binary variable C_{ij} is considered with $C_{ij} = 1$, if and only if paralog Q_j is assigned to contig X_i, and $C_{ij} = 0$ otherwise. Additionally, the following binary variable is introduced; E_{ijk}, with $E_{ijk} = 1$, if and only if TCE q_k^j from paralog Q_j is assigned to contig X_i, and $E_{ijk} = 0$ otherwise. While the variables C_{ij} represent the associations between paralogs and contigs, E_{ijk} represent the associations between the TCEs (of a certain paralog) and the contigs. The `ExonMatchSolver` then looks for an assignment that maximizes the total similarity score:

$$\max \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{k=1}^{m} \mu_{ij} \theta_{ijk} E_{ijk} \tag{2.2}$$

with θ_{ijk} being the bit score of the respective hit, and $\mu_{ij} = |\{k|\exists j' : \theta_{ij'k} > 0\}|$ being the number of (groups of homologous) TCE-hits found on contig X_i, i.e., those where for at least one paralog $Q_{j'}$ $\theta_{ij'k} > 0$. In addition to θ_{ijk}, which favors matches with a high similarity score, the factor μ_{ij} is introduced to prefer assignments with multiple TCE-hits found on the same contig.
The assignment is subject to a series of linear constraints. First, each TCE q_k^j is assigned at most once, and the same contig X_i does not carry more than one paralog Q_j.

$$\forall j, k : \sum_{i=1}^{n} E_{ijk} \leq 1 \quad \text{and} \quad \forall i : \sum_{j=1}^{N} C_{ij} \leq 1 \tag{2.3}$$

Second, a contig X_i is not assigned to paralog Q_j, if no TCE-hit q_k^j from paralog Q_j was found on this contig.

$$\forall i, j \text{ s.t. } \nexists k|\theta_{ijk} > 0 : \; C_{ij} = 0 \tag{2.4}$$

Third, contig X_i is assigned to paralog Q_j, if and only if at least one TCE q_k^j is assigned to that contig, i.e., $C_{ij} = 1$ if and only if $\exists k$ s.t. $E_{ijk} = 1$.

$$\forall i, j : \sum_{k=1}^{m} E_{ijk} - C_{ij} \geq 0 \tag{2.5}$$

$$\forall i, j : \sum_{k=1}^{m} E_{ijk} - mC_{ij} \leq 0 \tag{2.6}$$

with m being the number of groups of homologous TCEs.
Finally, if contig X_i is assigned to paralog Q_j, then all respective TCEs, which are found on this contig, are assigned to it, i.e., if $C_{ij} = 1$ then $\forall k$ s.t. $\exists j'$ for which $\theta_{ij'k} > 0$, it holds that $E_{ijk} = 1$. Otherwise, if $\forall j'\ \theta_{ij'k} \leq 0$, then $E_{ijk} = 0$.

$$\forall i, j : \mu_{ij} C_{ij} - \sum_{k|\theta_{ijk}>0} E_{ijk} \leq 0 \tag{2.7}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{k|\forall j':\theta_{ij'k}\leq 0} E_{ijk} = 0 \tag{2.8}$$

This simple ILP determines an optimal assignment C_{ij} of paralog Q_j to contig X_i, which can now be used to determine the sequences of paralogs. In these gene models, however, there still may be small or divergent exons missing, for which no significant hits were obtained.

2.2.4 Post-processing

To alleviate this limitation of the initial similarity search, two additional search steps are performed: (1) Local `tblastn` searches limited to only those contigs, where hits were identified for at least one TCE-model may identify additional candidate TCEs; (2) Spliced alignments of the query sequence on un-assembled contigs are used to increase the sensitivity. In contrast to local `tblastn` and `hmmsearch`, spliced alignment tools such as `ProSplign` can align the full-length protein query sequence to a genomic sequence fragment. This makes it possible to detect short TCEs that do not yield significant scores in genome-wide searches.
Upon compiling the final gene models, three cases appear: (1) In the simplest and ideal case, a paralog is located on a single contig with all TCEs fully covered and identified. No other assembly steps are required; (2) The paralog is distributed over multiple contigs such that every contig contains a sequence of consecutive TCE-hits in the correct order. In this case, the different fragments can be concatenated unambiguously, accounting for the TCE order and the strandedness of the fragments; (3) The TCE-hits identified on a contig are ordered correctly but they are not consecutive. For example, X_1 might carry TCEs $p...q$ and $r...s$, but $q+1...r-1$ are located on X_2. This occurs if the genome assembly is erroneous or if the two "contigs" are actually (pieces of) two scaffolds that interleave (e. g. Fig. 2.2, contigs a and h). To account for these cases, the pipeline attempts to insert X_2 in the appropriate place of X_1. The hypothesis of how two or more contigs have to be interleaved is entirely determined by the order of the exons on the query gene, and is therefore unique. If the contig contains stretches of Ns (indicating missing sequence at the scaffold level), the contig parts are interleaved there. Otherwise, the sequence is inserted at an arbitrary locus

for the preservation of the correct exon order. A spliced alignment tool is then run again on the merged contigs to refine the gene model.

2.2.5 Implementation and usage

The EMS-pipeline can be run in three general modes depending on the information available as input (indicated as different colors in Fig. 2.3). As the minimal input in all different modes, the protein sequences of all paralogs of interest from a well annotated species and the complete target genome must be provided as fasta files. In addition, the user can specify that the target genome has undergone a genome

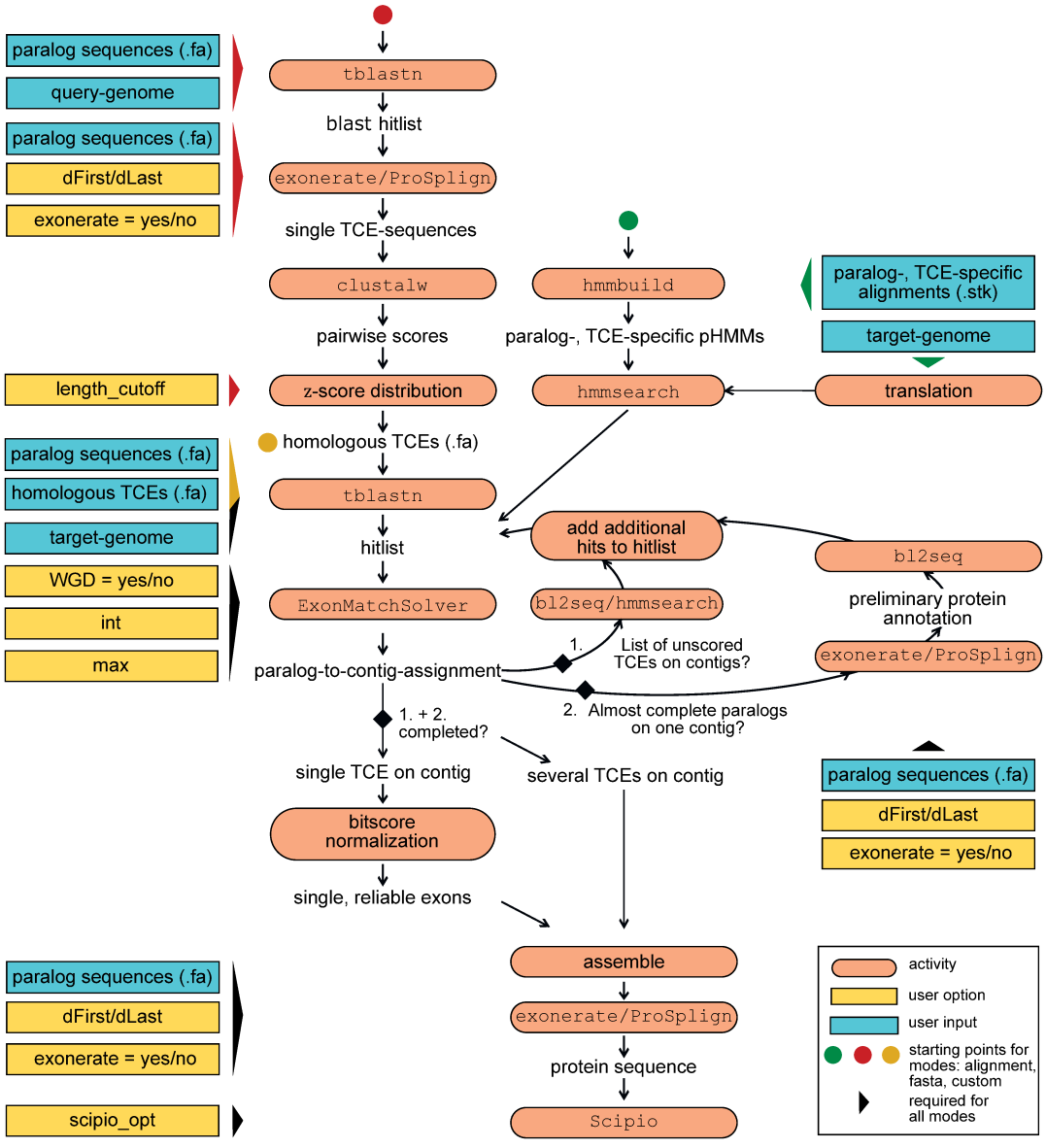

Figure 2.3: Extended schematic on steps, user-options and input-modes of the EMS-pipeline. The starting points for the three different modes of the EMS-pipeline are illustrated by red, yellow and green dots as are the different input files required for the respective mode. User options are given on the left and right side in yellow. See legend for further explanation. Abbreviations: pHMM – profile Hidden Markov Model; TCE – translated coding exon; WGD – whole genome duplication.

duplication relative to the query sequences, as is the case when using tetrapod queries to interrogate the genomes of teleosts (option `WGD`). In "fasta-mode" (red), homologous TCE groups are identified by a `tblastn` of the query protein against the query genome (Fig. 2.1, step 0b, Fig. 2.3). To reduce false assignments of TCEs to homologous groups, I compute a background distribution of pairwise similarity scores from the matches of a query TCE against all other TCEs of the same paralog. This information is used to determine a cut-off value $\hat{\theta}_j$ corresponding to a user-defined z-score to remove likely promiscuous matches between non-homologous TCEs. In order to further reduce the false assignments of short TCEs to homologous groups of putative lengthy TCEs, TCEs with lengths below a length cut-off are excluded (Fig. 2.3, option `length_cutoff`). This step may require manual inspection if exons are split or merged to increase the number of TCEs considered as input to step 1.
The "alignment-mode" (green) can be used when the exon–intron structure of the paralogs is already known and the user has access to well-annotated sequences from several species. Input protein alignments are converted to pHMMs applying the `HMMER3` suite (Eddy, 2011) and are then used to scan the conceptually translated target genome (Fig. 2.1, step 0a, Fig. 2.3). This improves both specificity and sensitivity of the tool. It can be used iteratively to improve results from a first set of searches starting from a single query.
Alternatively, the user can provide information on TCE-homology of the query protein sequences in "custom-mode" (yellow) to include as many homologous TCE groups as possible in comparison to "fasta" mode. The color coding of the different modes in Fig. 2.3 reflects how much information is provided by the user (green – most informative, red – least informative). Providing more information improves the performance of the EMS (green being the most sensitive).
Exact exon–intron structure of the query sequences in the target genome and in the query genome, if necessary, are inferred by means of a spliced alignment tool, by default `ProSplign` (Thibaud-Nissen et al., 2013). Alternatively, `exonerate` (Slater and Birney, 2005) can be used, which is faster but less sensitive (Hatje et al., 2011). In cases in which very long introns are predicted, the EMS-pipeline switches to `exonerate` automatically.
The ILP solver can be used to obtain alternative, suboptimal assignments (set option `max` to limit the number of returned solutions). This is particularly useful to judge the reliability of the solution. After completion of the first assignment by the `ExonMatchSolver`, the TCE-search is refined by running `hmmsearch` and `tblastn` with more sensitive settings as described above. The majority of TCE-hits for one paralog is usually assigned to one contig. A spliced alignment tool is used to align the query sequences to these contigs. The list of hits is augmented with these hits and the final paralog-to-contig assignment is computed.
Different contigs assigned to the same paralog are then merged/assembled. In some cases, contigs are interleaved. If so, the sequence of a single coding exon is inserted into the genomic area between the closest TCE-hits on the main fragment. If this region contains stretches of three or more consecutive Ns, the sequence is inserted in one of these regions. Large blocks of Ns are substituted by the insert-sequence. If the contig has no N-blocks in the appropriate region, the coding exon is inserted together with flanking Ns. The resulting edited "scaffolds" are again compared against the query sequences via a spliced alignment.
The resulting protein models as well as the input protein sequences are finally turned over to the `Scipio` gene annotation pipeline (Keller et al., 2008; Hatje et al., 2011). Gene annotations/hits proposed by `tblastn`, `exonerate`, `ProSplign` as well as `Scipio` can be compared by the user. The assignment list created by the

ExonMatchSolver and the list of any remaining, questionable, single coding exons is available for manual evaluation.

2.2.6 Assessment of the ExonMatchSolver's performance by simulations

In order to estimate performance and running time of the core step, I tested the ExonMatchSolver on simulated data. Protein sequence evolution is simulated with ALF (Dalquen et al., 2012) for two hypothetical species (query and target) in two steps. The first step allows for insertions, deletions, substitutions and duplications in a randomly generated protein sequence (branch length, $n = 50$, indel-rate = 0.0005, standard settings otherwise). It implements the evolution of one ancestor protein sequence to a fixed number of paralogs with an average of 2.5 % indels per sequence. The simulated protein sequences are divided into homologous pieces according to exon lengths sampled from a data set of human protein-coding genes originating from Ensembl (Lozada-Chávez, Stadler, and Prohaska, 2018). These exons are simulated to evolve independently (branch length, $n = 20$, about 1 % indels per sequence) without allowing for duplications in a second step representing recent evolutionary changes and accommodating rate differences within the protein. Exons of the single paralogs are distributed to different units (representing genomic fragments) with varying fragmentation levels. The fragmentation level is calculated as the average number of exons per fragment. Scoring of the query protein or TCEs against the target TCEs to identify homologous TCEs is performed with blastp (E-value $<$ 0.0001).
Performance of the ExonMatchSolver is assessed in comparison to a "greedy" method. A greedy assignment of a paralog to a unit is solely determined by the identity of the unit which retrieved the best bit score with the respective full-length query paralog. This best-hit approach is a very common strategy in gene annotation (section 1.3.4). Accuracy and running time of the ExonMatchSolver and the greedy method both depend on the individual random protein sequences that were simulated as well as on the exon sizes that are sampled from the exon length data set. To be able to directly compare these results, estimation of accuracy and running time are performed on the same set of simulated protein sequences. For the accuracy estimation, fragmentation is repeated 1,000 times for each fragmentation level with a fixed number of paralogs (8) and exons (12). The running time of the ExonMatchSolver is estimated for different numbers of exons and paralogs and a fixed fragmentation level (7.7 exons per fragment on average). The estimated user time is averaged over 20 different fragmentations on the same simulated data. Resident Set Size (rss) is used as an estimate of memory.

2.3 Results

2.3.1 Performance on simulated data

Accuracy of the ExonMatchSolver was estimated on simulated data and compared to the greedy method's accuracy on the same data set. For the simulated sequences of eight paralogs with 12 exons, the ExonMatchSolver solved the paralog-to-contig assignment more accurately than the greedy method if paralogs were fragmented across several units. For non-fragmented paralogs, the accuracy of the ExonMatchSolver was as good as that of the greedy method (Fig. 2.4 A). As expected, accuracy of both methods decreased with higher fragmentation of the genome, indicated by a lower number of exons per fragment. While the accuracy of the greedy method

dropped by more than 90 % from 1 to 0.08, the accuracy of the ExonMatchSolver solution did not fall below 0.91 even for the highest fragmentation levels. Thus, the ExonMatchSolver clearly outperformed the greedy method in assignment of paralogs to the correct units, which equalize contigs in non-simulated data.
In some simulations, the maximal accuracy of the ExonMatchSolver might be slightly lower than the accuracy of the greedy method at high fragmentation levels. This can be attributed to false negative hits representing short or very divergent exons that are not retrieved by the ExonMatchSolver. In the greedy comparison such false negatives do not occur because there, contigs are queried with the full-length protein. Although such false negative hits are in part retrieved in the post-processing step of the EMS-pipeline (as seen for the show case examples below), this step was not included in the performance tests for the ExonMatchSolver.
The running time of the ExonMatchSolver was in the range of a few seconds to minutes in dependence on the number of exons and paralogs (Fig. 2.4 B). Instances with 100 exons and 100 paralogs, the largest number of exons and paralogs tested, were an exception to this rule as they required about 2.5 hours of running time and 228 GB of memory on average. For more moderate numbers of 70 exons and up to 20 paralogs, running time was below one minute while at most 3.5 GB of memory were required. The running time and memory increased to more than 15 minutes and 35.4 GB, respectively, when exceeding 50 exons and 70 paralogs. The ExonMatchSolver thus solved even instances with extremely high numbers of paralogs and exons in feasible time. For most biologically relevant instances, memory requirements do not exceed the resources provided by a contemporary notebook.

2.3.2 Performance on real data - Two Showcase Examples

I selected two difficult examples, latrophilin receptors and arrestins to demonstrate the usefulness of the full EMS-pipeline on real data. Small differences in the exon–intron structure of the input paralogs are handled as if all paralogs derive from an ancestor that contains all coding exons.

Arrestins

As reported in section 1.5.4, four arrestin paralogs exist in human (*Homo sapiens*) that are encoded by 15-16 exons, *SAG*, *ARRB1*, *ARRB2* and *ARR3*. All arrestin genes except *ARRB1* are duplicated in zebrafish (*Danio rerio*) as a result of the 3R-WGD event (Renninger, Gesemann, and Neuhauss, 2011). The genes span a length of up to 82 kb. Overall, the exon–intron structure is conserved except for two intron losses in zebrafish *ARRB2b* and *ARR3a*. There are two micro-exons, exons 1 and 15, with less than 15 nt in length. These are particularly challenging to infer. I aimed to predict the seven arrestin paralogs in pufferfish (*Takifugu rubripes*, Ensembl *FUGU 4.0*) with the EMS-pipeline in "custom-mode" starting from protein sequences in zebrafish. If no experimentally verified entries were available in Genbank (NP_001153294.1, AAH76177.1, AAI52656.1, NP_957418.1), the annotations were extracted from Ensembl, *Zv9*. The last exon of *SAGb* was identified by an additional tblastn-search with *SAGa* as query resulting in a manually curated set of the input protein family. In the following, values for the number of contigs, to which paralogs were assigned, refer to the final output of the EMS-pipeline after spliced alignment of the assembled loci. TCEs were considered as found even if they were only partially identified. In the same sense, extensions of TCEs by the spliced alignment tools and additional alignment hits on the same fragment were not considered as false positives.

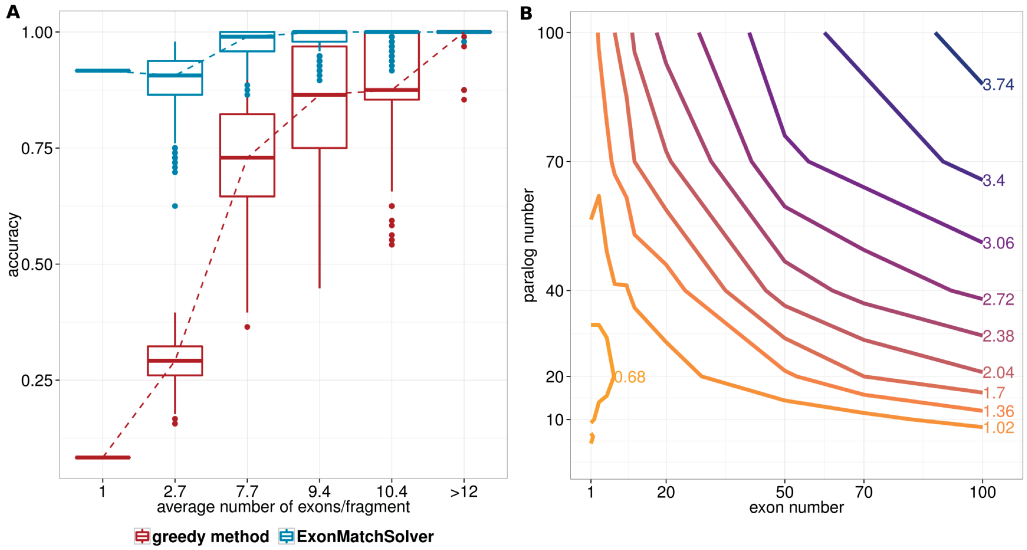

Figure 2.4: Accuracy and running time of the `ExonMatchSolver` on simulated data. A – Dependence of the accuracy on the fragmentation level in comparison with a greedy approach. Eight paralogs, each possessing 12 exons, were simulated in two species using ALF (section 2.2.6 for parameters settings). Fragmentation of exons across units was simulated 1,000 times for each fragmentation level. B – Dependence of the running time on paralog and exon number. Color changes of contour lines from yellow to dark blue indicate an increase in running time. Contour lines are labeled with the $\log_{10}$ of the running time. Different numbers of paralogs (4, 6, 8, 10, 20, 40, 70, 100) and exons (1, 3, 5, 7, 10, 12, 20, 50, 70, 100) were simulated using ALF with 7.7 exons per fragment on average. Running time was estimated as the mean of the user time of 20 runs with different fragmentation levels of the same simulated sequence data.

For the arrestins, the EMS-pipeline identified all expected seven arrestin paralogs situated on nine different contigs (Fig. 2.5 and Tab. 2.1). Five paralogs were (nearly) completely encoded on one contig each, while only parts of the other two, *SAGb* and *ARRB1*, were sequenced. *SAGb* and *ARRB1* were fragmented covering two genomic units each. For comparison, I ran `Scipio`, which identified four different arrestin loci with cross-species default options suggesting the loss of three paralogs relative to zebrafish. Considering the best scoring results for each query, `Scipio` assigned two different arrestin paralogs to *scaffold_525*, while three other paralogs were assigned to *scaffold_352*. No hits were suggested for *ARR3b*. Running `Scipio` with optimized options for arrestin genes allowed for an increased assembly size and increased sensitivity for detection of small exons. Therefore, seven different loci were proposed among the alternative results (see Fig. 2.6 for a phylogenetic tree of all alternative `Scipio` annotations). This is in accordance with the results proposed by the EMS-pipeline. Four out of the seven paralogs were correctly identified by `Scipio`, while the other three matched loci already assigned to a paralogous group (Tab. 2.1). In other words, none of the contigs harboring the other three orthologs *SAGb*, *ARRB1*, and *ARRB2b* appeared as a best scoring result in the `Scipio` predictions. In this example, the EMS-pipeline with `ProSplign` as spliced alignment component correctly identified two coding exons that remained undetected by `Scipio`. The short coding exon 1 of *ARR3a* (eight nt) could be annotated manually with the help of a local `blastn` search using the nucleotide sequence of the corresponding zebrafish exon as query. It was missed by both the EMS/`ProSplign`- pipeline and by `Scipio`.

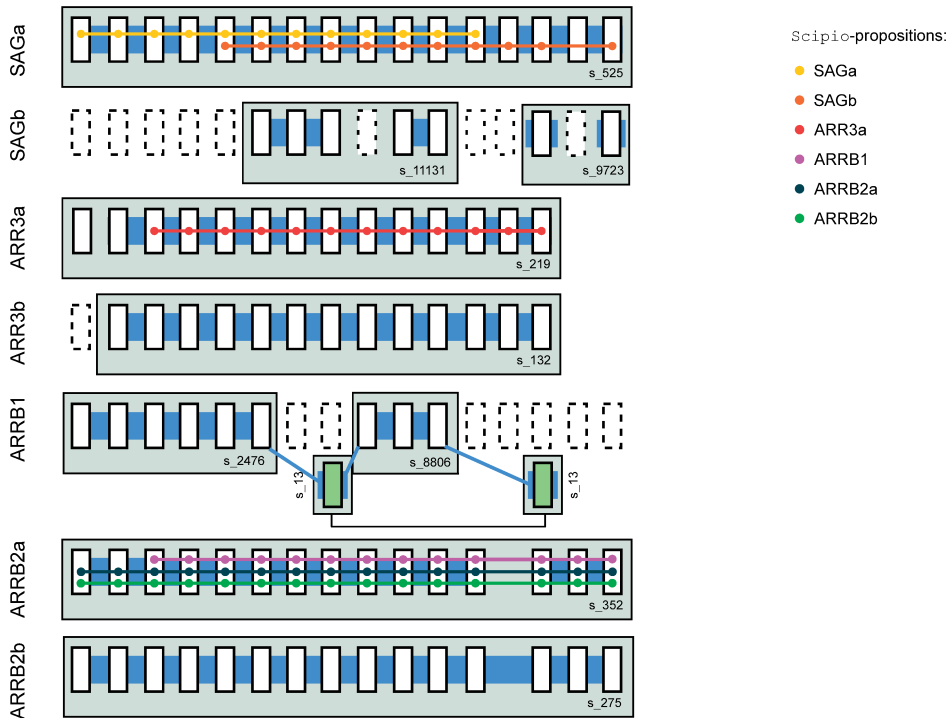

Figure 2.5: Illustration of the paralog-to-contig assignment for arrestin paralogs in pufferfish. All known zebrafish arrestins were used as queries. Homologous coding exons that were detected by the EMS-pipeline and `Scipio`, respectively, are shown as black open boxes on their respective contigs (grey boxes). Putatively missing (or deleted) coding exons are denoted by dotted boxes. False positive TCE-hits that were included in the `ExonMatchSolver`-solution, but not annotated by the spliced alignment tool are indicated by light green boxes. False positives appearing in the final output of either tool are indicated by brown boxes. The solution of the EMS-pipeline considering all of its stages is highlighted by broad blue paths in the back of the exons. `Scipio`'s best scoring proposition for each query is illustrated by colored dots and paths. Exon 16 is retrieved within the EMS solution, but not during the spliced alignment (*). Note that `Scipio` with cross-species default options did not propose any hit for *ARR3b*. Abbreviations: GS – GeneScaffold; c – contig; s – scaffold; TCE – translated coding exon.

This inference is also an example for an instance, where the phylogenetic tree does not easily resolve the orthology relationship. While pufferfish *ARR3a*, *ARRB1*, *SAGa* and *ARRB2a* form a monophyletic group with their respective zebrafish orthologs, this is not true for the other three paralogs (Fig. 2.6 A, B).

Latrophilins

The latrophilins (*ADGRL1*, *ADGRL2* and *ADGRL3*) belong to the family of adhesion G protein-coupled receptors (GPCRs) and are encoded by 22-26 exons spanning a total length of up to 210 kb in zebrafish (Silva and Ushkaryov, 2011). A recent phylogenetic study proposed the duplication of *ADGRL1* and *ADGRL2* in zebrafish resulting in a greatly shortened N-terminus (Harty et al., 2015). The five paralogous family members have a highly similar exon–intron structure in zebrafish thus fitting well with the application scenario of the EMS-pipeline. In *ADGRL1a* and *ADGRL1b*, exon 5 is split into three independent exons in comparison to the other paralogs, resulting in 25 homologous exon groups. I aimed to annotate *ADGRL1*, *ADGRL2* and

Figure 2.6: Phylogenetic tree of arrestins as annotated by `Scipio` in pufferfish together with query orthologs from zebrafish. `Scipio` was run in sensitive mode (modified options: `max_assemble_size=50000`, `min_score=0.1`, `exhaust_align_size=50000`, `region_size=90000`). The alignment and neighbor joining-tree on protein level were built with `Clustal Omega 1.2.1` (Sievers et al., 2011) on all columns of the alignment (A) and on all columns that did not contain gaps (B). The alternative solutions of `Scipio` on the same genomic unit may slightly differ in dependence on the query paralog it was retrieved with (indicated by the first part of the node label). The zebrafish paralogs are denoted as *SAGa,b*, *ARRB1*, *ARRB2a,b* and *ARR3a,b*. The solution proposed by the EMS-pipeline is highlighted in green. *ARRB1* is situated on `scaffold_2476_scaffold_8066`. Due to long branch attraction, missing data and sequence divergence, zebrafish and pufferfish orthologs do not always form monophyletic groups. This makes inference of paralog-to-contig assignments based on the phylogenetic tree difficult. The trees were edited and displayed with `Dendroscope 2.6.1` (Huson et al., 2007).

ADGRL3 and possible additional paralogs in cod (*Gadus morhua*, `Ensembl` *gadMor1*), which shares the 3R-WGD with zebrafish. As a starting point, I chose the annotation of latrophilin paralogs in the well assembled genome of zebrafish (`Ensembl` *GRCz10*).

Table 2.1: Performance of `Scipio` and the EMS-pipeline in prediction of arrestin genes in pufferfish. The pufferfish genome *FUGU 4.0* (`Ensembl`) was queried with zebrafish protein sequences (NP_001153294.1, AAH76177.1, AAI52656.1, NP_957418.1 and annotations from `Ensembl` *Zv9*). `Scipio` was run with "cross-species default options" (`min_identity=60, max_move_exon=6, blat_score=15, blat_identity=54, multiple_results, region_size=10000, exhaust_align_size=15000`, results given in bold) and in a more sensitive mode (modified options: `max_assemble_size=50000, min_score=0.1, exhaust_align_size=50000, region_size=90000`). The sensitive `Scipio`-mode included all hits of the cross-species default options. If scores deviated, these are separated by "/". As `Scipio` was run with the multiple_results-option, several hits are occasionally returned; these are indicated by a number in brackets in the paralog-column. The EMS-pipeline was run in "custom-mode" with `ProSplign` as spliced alignment tool. TCE-numbering refers to the homologous TCE-groups. The EMS-pipeline returned correct contig assignment for paralogs, which were predicted to be situated on different contigs by `Scipio`'s first hit (marked in red). Hits were considered even if they were partial only. Abbreviations: fp – false positive; s – scaffold.

`Scipio`				EMS-pipeline		
paralog	contig	score	TCEs identified	contig assignment	TCEs included by the `ExonMatchSolver`	TCEs included after post-processing (`ProSplign`)
SAGa	s_525	0.426	1-12	s_525	3-12, 14, 16	1-16
SAGb	s_525	0.322	5-14, 16	s_11131	6-8, 10-11	6-8, 10, 11
SAGb(1)	s_275	0.151	3-7, 11, 12, 14	s_9723	14, 16	14, 16
SAGb(2)	s_11131	0.127	6-8, 10, 11			
	s_9723	0.052	14, 16			
ARRB1	s_352	**0.536**-/0.538	3-12, 14-16	s_2476	2-6	1-6
***ARRB1*(1)**	s_275	0.44	2-14	s_8806	9-11	9-11
ARRB1(2)	s_2476	0.225	2-6	s_13	8, 14 (fp)	-
	s_8806	0.187	9-11			
ARRB2a	s_352	1.000	1-12, 14-16	s_352	2-12, 14-16	1-12, 14-16
***ARRB2a*(1)**	s_275	0.527	2-12, 14			
ARRB2a(2)	s_2476	0.167	2-6			
	s_8806	0.150	9-11			
ARRB2a(3)	s_525	0.126	3, 8-12			
ARRB2b	s_352	**0.797**-/0.819	1-12, 14-16	s_275	3-12, 14, 16	1-12, 14-16
***ARRB2b*(1)**	s_275	0.529	2-12, 14			
ARRB2b(2)	s_2476	0.162	2-6			
	s_8806	0.154	9-11			
ARRB2b(3)	s_525	0.152	3, 5-12			
ARR3a	s_219	0.457	3-14	s_219	5-12, 14	2-14
ARR3a(1)	s_132	0.367	2-14			
ARR3b	s_132	0.238	2-11	s_132	3-12, 14	2-14
ARR3b(1)	s_219	0.210	5-12			

To obtain a trustworthy query, these were manually curated adding small, missing exons identified by tblastn with human latrophilins as queries. During curation, an additional paralog, *ADGRL1b*, was identified. The starting data set thus comprised six latrophilin paralogs, all of which were also identified in cod with the EMS-pipeline. Presumably due to missing data, in total, the sequence of five different single TCEs was missing for all latrophilin paralogs in the zebrafish query in total, which causes the spliced alignment tool missing exactly those exons. As a byproduct, the ExonMatchSolver will keep TCE-hits even if the fragment is scored with a different paralog-model of the same TCE only. This results in detection of coding exons that might either be missing from the query paralog or represent a pseudogenic exon (marked by asterisk, Fig. 2.7).

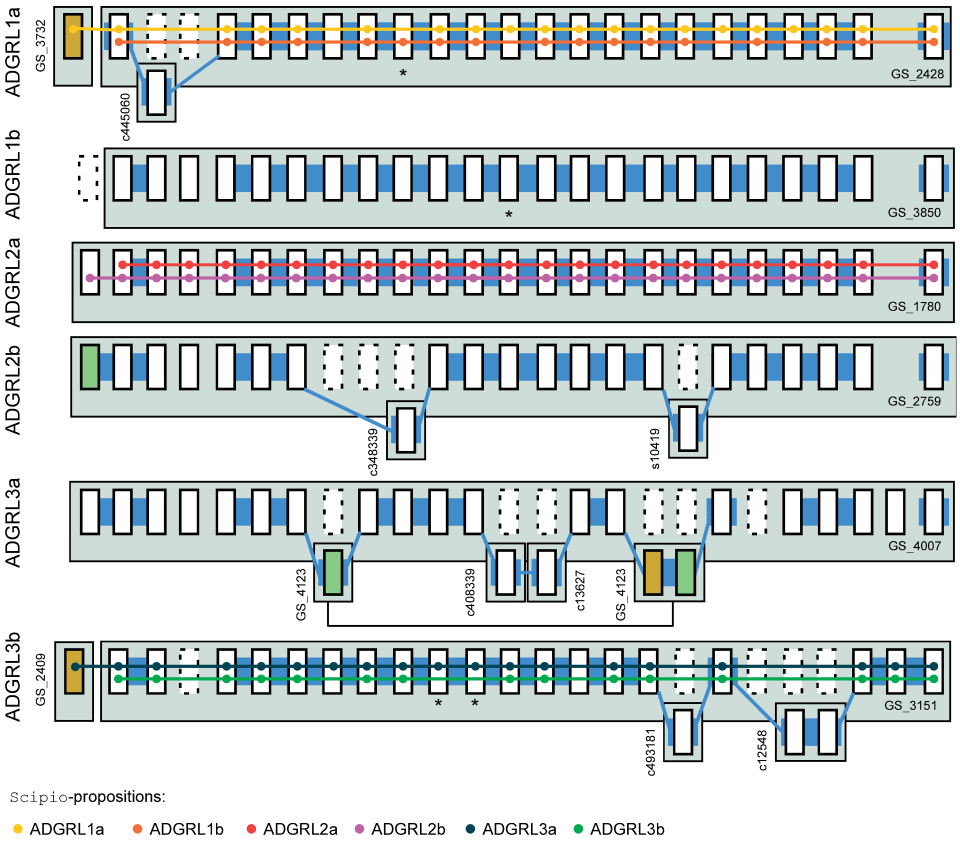

Figure 2.7: Illustration of the paralog-to-contig assignment for latrophilin paralogs in cod. All known zebrafish latrophilins were used as queries. Please see Fig. 2.5 for a description of colors/symbols. True positive TCEs that were detected by the ExonMatchSolver, but not by the spliced alignment tool, are marked with an asterisk (*). Abbreviations: GS – GeneScaffold; c – contig; s – scaffold; TCE – translated coding exon.

The EMS-pipeline identified all six paralogs existing in zebrafish situated on 14 different fragments in cod. In contrast, Scipio (Hatje et al., 2011) placed the latrophilin paralogs onto five different contigs or scaffolds in cod when run under cross-species default options (Fig. 2.7, Tab. 2.2). Considering the best scoring results only, the tool proposed the existence of three different latrophilin loci. At these loci, Scipio proposed each of the recently duplicated paralog-pairs shared the exact same coordinates on one fragment. If instead, the user inspected the alternative results for each paralog, Scipio's next-best scoring fragments did not necessarily correlate

with the correct contigs that were found by the EMS-pipeline. This was the case for exon 1 of *ADGRL1a* and *ADGRL3b*, which could be identified as false negative hits by manual inspection. The EMS-pipeline instead suggested eight different contigs to be interleaved with four of the main fragments. Eight of these nine TCE-hits, proposed in the final output, likely represent true exons that were situated on short fragments remaining from an incomplete genome assembly. In the available annotation of cod, no further genes were annotated on these fragments, supporting the correct paralog-to-contig assignment.
The ninth hit corresponds to exon 23 of the gene *CELSR1b* encoding part of a secretin-like domain thus representing a false positive hit of the EMS-pipeline. Exons 15-20 of the latrophilin genes code for this domain, common to the whole class of adhesion GPCRs. Inspection of the initial `tblastn`-hitlist retrieved several high scoring hits of more distant paralogs (e. g. *ADGRL4*, *ADGRE5*, and unnamed genes with GPCR-domains) that all possess this domain.
The use of `exonerate` as a spliced alignment tool caused the EMS-pipeline to miss the short exon 4 in all latrophilin paralogs (15 nt), the short exon 24 in *ADGRL3a* (18 nt), and the divergent exon 1 of *ADGRL2a* that were identified by `Scipio` in the alternative propositions. Therefore the usage of `ProSplign` with the EMS-pipeline is recommended whenever sufficient computational resources are available. Furthermore, the results of `Scipio` that are additionally returned by the EMS-pipeline can provide further improvement but require manual inspection.
Interestingly, in both, cod and zebrafish *ADGRL2b* and *ADGRL1b*, the exon–intron structure and overall protein length were conserved relative to *ADGRL2a* and *ADGRL1a*. This contradicts the proposed truncation of these two genes reported in Harty et al. (2015) and emphasizes the need to manually curate database annotation carefully considering differences in gene structure of paralogous genes.

2.4 Discussion

Applying a decomposition of proteins into TCEs and separation of homologs into their paralogous groups allows the EMS-pipeline to build models for individual paralog-specific TCEs. Combining the strengths of different well-established methods and tools (`ProSplign`, `exonerate`, `tblastn`, `HMMER` and `Scipio`) that translate between the level of protein and genomic sequence, and novel algorithmic approaches (the automated paralog-to-contig assignment), the EMS-pipeline provides a comprehensive and flexible toolbox for manual, high-quality curation of gene annotations.
The core of the pipeline is the ILP formulation of the PCAP referred to as `ExonMatchSolver`, which is NP-complete. The `ExonMatchSolver` solves the assignment problem within seconds or minutes for most biologically relevant numbers of paralogs and exons in simulations. Even for high numbers of paralogs, which might occur in polyploid species such as the octaploid sugar cane (Setta et al., 2014), the running time does not exceed one hour for up to 70 exons. However, genes with more than 70 exons are rare for human and most other animals (Scherer, 2010).
The EMS-pipeline helps to overcome many of the critical problems arising from highly fragmented draft genome assemblies as demonstrated with simulated data as well as with two real life examples. The only program that has been targeted to solve a similar problem with the focus on a single gene family is, to my knowledge, `Scipio`. As suggested by one reviewer, one could alternatively use a maximum weight bipartite matching to identify the correct paralog-to-contig assignment among alternative `Scipio` solutions or build a phylogenetic tree of all alternative `Scipio` annotations

Table 2.2: Performance of Scipio and the EMS-pipeline in prediction of latrophilin genes in cod. The cod genome *gadMor1* (Ensembl) was queried with zebrafish protein sequences (Ensembl *GRCz10*). Scipio was run with "cross-species default options" (min_identity=60, max_move_exon=6, blat_score=15, blat_identity=54, multiple_results, region_size=10000, exhaust_align_size=15000). Scipio occasionally returned several hits; these are indicated by a number in brackets in the paralog-column. The EMS-pipeline was run in "custom-mode" with exonerate as spliced alignment tool. The TCE-numbering refers to the homologous TCE-groups. Hits were considered even if they were partial only. The EMS-pipeline returned correct contig assignment for paralogs, which were predicted to be situated on different contigs by Scipio's first hit (marked in red). Abbreviations: c – contig; fp – false positive; GS – GeneScaffold; s – scaffold.

Scipio				EMS-pipeline		
paralog	contig	score	TCEs identified	contig assignment	TCEs included by the ExonMatchSolver	TCEs included after post-processing (exonerate)
ADGRL1a	GS_3732 GS_2428	0.005 0.681	1 (fp) 2, 5-7, 9, 11-23, 25	GS_2428 c445060	2, 5-7, 9-23, 25 3	2, 5-9, 11-23, 25 3
ADGRL1a(1)	GS_1780	0.354	2-7, 9, 11, 12, 14-20, 22, 23, 25			
ADGRL1b	GS_2428	0.411	2, 5-7, 9-12, 14-18, 20, 22, 23, 25	GS_3850	2, 3, 5-20, 22, 23	2, 3, 5-12, 14-23, 25
ADGRL1b(1)	GS_3850	0.309	2-12, 14, 15, 17, 22			
ADGRL2a	GS_1780	0.643	2-23, 25	GS_1780	2, 3, 5-23, 25	2, 3, 5-23, 25
ADGRL2a(1)	GS_2759 GS_1138	0.425 0.005	2-7, 11-17, 19-20, 22, 23, 25 25			
ADGRL2a(2)	GS_871 GS_2428	0.005 0.316	1 2, 5-7, 9-12, 14-18, 20-23, 25			
ADGRL2b	GS_1780	0.527	1-23, 25	GS_2759 s10419 c348339	1 (fp), 2-3, 5-7, 11-17, 19-23, 25 18 10	2, 3, 5-7, 11-17, 19-23, 25 18 10
ADGRL2b(1)	GS_2759	0.513	2-7, 11-17, 19-23, 25			
ADGRL2b(2)	GS_3429 GS_2428	0.009 0.300	1 2, 5-7, 9-18, 20, 22, 25			
ADGRL3a	GS_2409 GS_3151	0.007 0.463	1 (fp) 2, 3, 5-14, 16, 17, 19, 23-25	GS_4007 GS_4123 c408339 c136327	2, 3, 5-7, 9-12, 15, 16, 19, 21-23, 25 8 (fp), 17, 18 (fp) 13 14	1-3, 5-7, 9-12, 15, 16, 19, 21-23, 25 17 (fp, belongs to *CELSR1b*) 13 14
ADGRL3a(1)	GS_4007	0.453	1-7, 9-12, 15, 16, 19, 21-25			
ADGRL3b	GS_3151	0.533	2, 3, 5-7, 9, 10, 13, 14, 16, 17, 19, 23-25	GS_3151 c12548 c493181	2, 3, 5-12, 15-17, 19, 23, 25 21, 22 18	2, 3, 5-10, 13-17, 19, 23-25 21, 22 18
ADGRL3b(1)	GS_4007	0.402	2-7, 9, 10, 15, 16, 19, 21-23, 25			

together with the query paralog sequences. As demonstrated in this example (Fig. 2.6) phylogenetic trees are often far from easy to interpret and may also require manual inspection for identification of the correct paralog-to-contig assignment. In cases, in which `Scipio` did not find the correct combination of genomic units for a paralog as for the latrophilin example, the problem may aggravate. The EMS-pipeline is designed to specifically fill this gap for detailed exploration of the evolution of a specific gene family of interest. The explicit use of exon–intron structures and the exon-centric computation of protein similarities furthermore improves the accuracy of paralog identification.

Given the diverse sources of errors and exceptional cases, I have not attempted to construct a fully automatic pipeline, but rather a tool to assist in manual data curation. As a similarity-based method, it depends heavily on the availability of high quality protein sequences (or alignments) as input queries. Erroneous exon annotations or splice site predictions leading to erroneous translated coding sequences in the input unavoidably will be carried over to the results and cannot easily be identified by automatic means.

At present, there are no databases that simultaneously provide both, paralogy information and accurate information on exon–intron structure. The exon–intron database (`EID`, Shepelev and Fedorov (2006)) and `SpliceDB` (Burset, Seledtsov, and Solovyev, 2001) do not provide information on paralogs; `Ensembl Compara` on the other hand, does not provide homology information for individual exons. The lack of a gold standard makes it unfeasible to quantitatively benchmark the EMS-pipeline on real data. Therefore, I demonstrated the superior accuracy of the `ExonMatchSolver` in comparison with a greedy method on simulated data. On real data, I had to rely on a few difficult use cases for which a detailed manual curation was possible.

In its present state, the EMS-pipeline has several limitations. Most importantly, I assume a largely conserved exon–intron structure of the paralogs, a situation that is very often encountered for vertebrate genes. Nonetheless, the exon–intron structure of distant relatives may differ strongly. This may limit the application of the EMS-pipeline to deuterostomes or clades within protostomes that conserve the gene structure in the gene of interest. Largely distinct gene structures can also be accommodated by treating the respective genes as separate paralogous groups. However, cases of recognizable structural similarity together with changing variability might be difficult to handle. Furthermore, I assume that a fairly complete collection of paralogs is used as an input. The paralog-to-contig assignment step may yield incorrect results if the *a priori* estimate of the number of paralogs is incorrect (as in the latrophilin example). In particular, this may lead to the inclusion of more distant, spurious solutions or result in fragmented gene models. In these cases, manual inspection of the results thus appears unavoidable. I therefore have designed the EMS-pipeline to streamline and simplify the process of manual post-processing that is required for most fragmented genes.

Several improvements in future releases of the EMS-pipeline are planned in response to exceptional cases that were encountered in practical tests so far: The number of paralogs in a genome can presumably be estimated by a more careful analysis of the spectrum of similarity scores (Chapter 4). This should help to largely prevent the inclusion of false positives to "compensate" for lineage-specific gene losses and would be useful also when studying gene families with many levels of paralogy, i.e., large numbers of nested gene duplications. It may also be possible to improve the accuracy of the initial, score-based assignments of coding exons to paralogs by using a reciprocal best hit strategy rather than relying on the bit scores θ_{ijk} of the query matches alone.

Besides Scipio, the EMS-pipeline is, to my knowledge, the only gene-focused toolkit that can deal with the fragmentation of genes across different contigs in a systematic manner. With their SWiPS pipeline, Li and Copley (2013) provide a similar approach to the ExonMatchSolver, although they set a different focus: the improvement of a complete genome assembly with the help of protein annotations. The ExonMatchSolver, instead aims to find the best solution considering a single gene family, which is connected to a substantially lower computational effort.

Chapter 3

Evolution of the arrestin protein family in deuterostomes

This Chapter is based on Indrischek et al. (2017). The respective protein residue numbering refers to the bovine ortholog (cow, *Bos taurus*) unless stated otherwise. The exon–intron structure naming is based on homology and consistent across different orthology groups (see Fig. 3.16 A for reference).

3.1 Motivation

Arrestins are cytosolic signaling transducers that directly bind to activated and phosphorylated G protein-coupled receptors (GPCRs, explained in the introduction, section 1.5). They constitute early key players of different signaling cascades as they mediate receptor desensitization via competition with G proteins (section 1.5.1). As signaling cascades regulate key cellular processes such as cell replication and apoptosis, the arrestin protein family is an attractive therapeutic target. The signaling outcome is believed to be a result of structural and sequence-dependent interactions of the activated arrestin with GPCRs and other protein partners (section 1.5.2). Besides their role in blockage of G protein activation, arrestins are scaffolding hubs that form the physical link between post-translational modifying enzymes and their substrates. Furthermore, non-visual arrestins mediate internalization of GPCRs by interacting with both GPCRs and the endocytosis machinery (section 1.5.3). The arrestin family interacts with numerous interaction partners despite consisting of just four family members in mammals. A deeper understanding of these interactions and their resulting downstream effects is a necessity for a future selective regulation of these processes by drugs.

A popular and often visited route from classical biochemistry answers questions about the sequence and structure–function relationship by performing point mutations of positions of interest and evaluating their functional and structural effects *in vitro* or *in vivo*. Thinking of this problem from a different direction, one can argue that all of those mutations and many more have already been tested during evolution and can be studied by investigating the evolutionary history of arrestins. Patterns of conservation, covariation and selection can reveal properties about interaction interfaces. Neo- and subfunctionalization might reveal how existing functions can be modified or re-used in a different context. Gene deletions on the other hand might provide hints on which functions might be redundant.

While the cloning of individual arrestins has already led to the discovery of unexpected duplications and subfunctionalizations, the evolutionary history of arrestins has not been studied systematically (section 1.5.4). The information on arrestin homologs presented in literature either covers only a very limited range of species (Alvarez,

2008) or an incomplete set of paralogs for most species investigated (Gurevich and Gurevich, 2006a). On the other hand, homology search solely based on domains lacks resolution on exact orthology relationships (Mendoza, Sebé-Pedrós, and Ruiz-Trillo, 2014). For this reason, the objective of this Chapter is to systematically investigate the evolutionary history of arrestins in vertebrates and their close relatives (deuterostomes). I first provide an overview of the arrestin fold family in animals (Metazoa) and beyond, which contains arrestins and other proteins, which have *arrestin_N* and *arrestin_C* domains by querying protein databases. After defining the group of interest, arrestins, this family is investigated in detail based on a re-annotation in deuterostome genomes with the `ExonMatchSolver`-pipeline (EMS-pipeline, Chapter 2). This Chapter is focused on arrestins in deuterostomes as hypothesis about sequence–function relationships are restricted to 1:1 orthologs, which are well studied in mammals in the case of arrestins. Sequence and exon–intron structure conservation are evaluated to gain insights into possible functional changes of the less studied members of the protein family and to elucidate nature's repertoire of signaling interfaces relating to arrestins.

3.2 Material and Methods

For more background information on tools employed, please refer to sections 1.3, 1.4 and Chapter 2.

3.2.1 Database scan

For performing a homology search of arrestins against the `UniProtKB` database (accessed via `https://www.ebi.ac.uk/Tools/hmmer/`, February 2017), I generated a profile Hidden Markov Model (pHMM) using `jackhmmer` with an alignment of the four human arrestins as input (*Homo sapiens*, Finn et al. (2015)). Running `jackhmmer` for a higher number of iterations will retrieve more distant homologs. For each iteration, a new pHMM is built from the homologs retrieved in the previous iteration. For searching homologs of the arrestin family, the number of `jackhmmer` iterations was chosen so that the `jackhmmer` set of homologs showed a good overlap with the results of a homology search with the pHMMs of the domains *arrestin_N* and *arrestin_C* as downloaded from `Pfam 31.0` (PF00339, PF02752, E-value < 1). The full-length set of homologs obtained from `UniProtKB` was filtered according to length ($422 >$ length > 195, μ +- σ), E-value (< 1) and identity of the full-length sequences for each species separately (< 80 %). The identity filter cut-off was chosen to balance the removal of isoforms and retention of paralogs and to contain the expected number of paralogs for human, cow and fruit fly (*Drosophila melanogaster*). An identity cut-off of 85 % retrieved a false-positive isoform/paralog for human, while an identity cut-off of 90 % discarded a true-positive isoform in stickleback (*Gasterosteus aculeatus*). I obtained a set of 2962 sequences, 2348 of which contained at least one *arrestin_N* and one *arrestin_C* domain (Fig. A.1). 142 sequences did not have either of both domains and were excluded. I proceeded with the full-length sequences of this set under exclusion of hits that were not assigned to one specific species but to a clade (e. g. bilaterians), for phylogenetic inference, and for reporting paralog counts projected on the `NCBI` phylogeny.
In order to exclude effects on phylogenetic inference that can arise from aligning sequences that are not homologous in full length, I additionally generated individual domain sets for the *arrestin_N* and *arrestin_C* domain, separately, and also proceeded

to phylogenetic inference. These sets consist of the respective Pfam model hit in the UniProtKB database restricted to the actual hmmsearch hit length. Both sets were filtered according to identity (see above). As a consequence, sequences of proteins that contain more than one specific arrestin domain are contained several times within the alignment and respective tree.
Furthermore, I queried OrthoDB (as of February 2017) with the same full-length arrestin pHMMs (E-value < 0) obtained with jackhmmer. OrthoDB is considered to be a high quality orthology database, which contains unique orthology group assignments for proteins of interest on a given taxonomic level. I restricted the analysis to the OrthoDB groups that are annotated on animal level, which is the highest/most inclusive level for arrestins. Applying this strategy, 3487 hits were retrieved that belong to 109 orthology groups. For better visibility, only groups with more than 29 members are distinguished for plotting the results. These nine groups cover 88 % of all sequences. The NCBI species tree was retrieved with the ete toolkit (Huerta-Cepas, Serra, and Bork, 2016).

3.2.2 Detailed gene annotation

Automated methods frequently fail to correctly predict multi-exon genes (section 1.3.1, Chapter 2). I therefore used exon- and paralog-specific pHMMs implemented within the EMS-pipeline to update the annotation of arrestin genes in different genomes of interest. Exon models were built from an initial, manually curated protein alignment of mammalian arrestins. In order to generate this alignment, human arrestin reference sequences were retrieved from UniProtKB. These correspond to the well characterized and on transcriptome level supported annotations of the longest isoforms of three of the four arrestin paralogs in human annotated by Ensembl (Flicek et al. (2014), see Tab. B.2 for an overview of all isoforms).
First, annotation of arrestin homologs in 13 different mammalian orders were systematically completed. To do so, query protein sequences were blasted against the respective genome of interest using tblastn on the Ensembl web interface (Altschul et al., 1990). Missing short exons were retrieved using local tblastn or blastn (bl2seq 2.2.26, E-value < 1) and the spliced alignment tool ProSplign (Thibaud-Nissen et al., 2013). The reference sequence for *ARRB2* (409 aa) does not contain the 22 aa extension of exon 5 seen in the longest isoform in human (Flicek et al. (2014), Tab. B.2). The human isoforms chosen initially are homologous to each other in full length apart from minor deviations in the exon–intron structure and thus satisfy the requirements for application of the EMS-pipeline.
Second, an initial alignment was built from these sequences. The exon- and paralog-specific protein alignment of mammalian arrestins was then extended by adding the Translated Coding Exon (TCE) sequences from arrestins successively annotated in other clades. pHMMMs were built with HMMER 3.1b1 (Eddy, 2011), which was called in "alignment-mode" of the EMS-pipeline (Indrischek et al. (2016), Chapter 2). In case of a systematic failure to detect a specific arrestin exon within a monophyletic family with the EMS-pipeline, the candidate region was re-investigated with different homology-based methods. These included querying a region between two exon hits with local blastall 2.2.26, using as query the nucleotide sequence of the missing exon(s) (blastn), or the amino acid sequence of the conceptually translated missing exon(s) (tblastn), respectively (Altschul et al., 1990). To detect exons that differed substantially among homologous groups and could not be detected with any other method, the corresponding regions of at least three close relatives of one group were aligned with tba.v12 (Blanchette et al., 2004). The conservation-based

RNAcode 0.3 method was applied to detect conserved regions with protein-coding potential (Washietl et al., 2011). This strategy was e. g. applied to sauropsid *SAG* and *ARR0* exon 1 and teleost *ARR3* exon 16. In both cases, no TCE sequence was retrieved that was systematically conserved in the respective monophyletic group.

3.2.3 Genomes, transcriptome, Expressed Sequence Tag and Short Read Archive data used in the current study

Unmasked genomes were extracted from Ensembl, EnsemblPre! or Ensembl Metazoa if available and from the listed sources otherwise (Tab. B.1). For ghost shark (*Callorhinchus milii*), only a soft-masked version of the genome was available. To clarify the potential loss of *ARRB2* in birds, all available 48 bird genomes from the Avian Phylogenomics Project (Zhang et al., 2014) as well as the genomes of kiwi (*Apteryx australis mantelli*) and golden eagle (*Aquila chrysaetos*) were investigated additionally. All four arrestin paralogs were annotated in nine birds in total (ostrich, chicken, turkey, duck, finch, ibis, hoatzin, cuckoo, bald eagle). Insertions and stop codons were occasionally observed within exons of arrestin genes in genomes with low coverage and/or poor quality assemblies. Those were interpreted as sequencing or assembly errors because the remainder of the protein-coding sequence was usually highly conserved, except in cases which were explicitly identified as pseudogenes in the current study (e. g. elephant *ARR3*). Sequencing errors might also effect the protein-coding sequence of arrestin genes in those low quality genome. They cannot be distinguished from substitutions in the current study.
Transcriptome data sets, in particular the NCBI Expressed Sequence Tag (EST) and NCBI Transcriptome Shotgun Assembly data sets, were additionally queried whenever the analysis of the corresponding genome was not conclusive in regard to the presence and absence of gene copies. The NCBI webinterface was used to tblastn with protein sequences of closely related species as queries in these cases (Tab. B.3). Clades that were queried are "Sauropsida", "Aves", "Marsupilia", "Chondrichthyes" and "Cyclostomata" (National Institute of Health (US) and National Center for Biotechnology Information (*Translated BLAST: tblastn*) as of November 2015). NCBI Short Read Archive (SRA) was queried with the known arrestin kiwi exons against SRA data of ostrich (*Struthio camelus*) and tinamu (*Tinamus guttatus*) as well as with arrestin exons from bald eagle (*Haliaeetus leucocephalus*) against SRA data of white-tailed eagle (*Haliaeetus albicilla*) and golden eagle. As the NCBI BLAST did not provide a BLAST database for EST data of lizard (*Anolis carolinensis*), this was locally built and queried.

3.2.4 Alignment and building of phylogenetic trees

For generating a bootstrapped phylogenetic tree of the arrestin fold family, I aligned all hits obtained after filtering from the OrthoDB with Clustal Omega 1.2.4. Next, I built an approximate maximum likelihood (ML) tree with FastTree (Price, Dehal, and Arkin, 2010) with the -pseudo option for fragmented/gapped sequences and the following options to increase its accuracy/tree exploration -spr 4 -mlacc 2 -slownni (section 1.4.1).
For the tree of arrestins, I considered Genbank annotations of arrestins with experimental evidence (NP-entries) whenever they were available and more complete in regard to coverage than the genomic annotations retrieved in this study. The same holds true for transcript evidence of arrestin paralogs. Coding DNA sequences were aligned according to codons with MACSE 1.01b (Ranwez et al., 2011) and further

edited in mega 4.0.2 (Tamura et al. (2007), section 1.4.2). ML trees were built from protein sequences using PhyML 3.0.1 (Guindon et al., 2010). Optimal model parameters were determined using ProtTest 3.4 allowing for the following substitution models: JTT, PAM250, WAG, LG, DCMut, BLOSUM62, an estimation of amino acid frequencies (-F), a fraction of invariable sites and a gamma-distribution (-all-distributions, Darriba et al. (2011), section 1.4.1). Unknown amino acids were substituted by "?" in the alignment for tree building. The tree that obtained the best information content (Bayes Information content, BIC and Akaike Information content, AIC) applying ProtTest was used as starting tree for PhyML. The tree topology was validated by bootstrapping (1,000 iterations unless stated otherwise). Manual inspection of the alignment revealed conservation or disruption of functional motifs previously investigated experimentally in mammals and known from literature, that were marked within the Jalview 2.8.2 alignment program (Waterhouse et al., 2009).
Bayesian trees were constructed based on the amino acid alignment with the BEAST2 software (Bouckaert et al., 2014) under the birth–death model with a relaxed molecular clock (section 1.4.1). JModelTest v.2 was used to test for substitution models of nucleotide alignments, which were set as prior parameters in Beauti/BEAST (section 1.4.1). I compared different model settings pairwise by employing PathSampling (Baele et al., 2012; Baele and Lemey, 2013) to estimate the marginal likelihoods and to calculate the Bayes factor (BF). The model settings differed in their birth–death priors and regarding estimation or fixation of different priors to specific values, while using the parameters determined with ProtTest as site model parameters for the amino acid alignments/trees. Models were excluded if they yielded infinite likelihood estimates or did not converge (see Tab. B.4 for parameters and best models). As the unconstrained gene trees did not have the expected topology with *ARR0* as an outgroup of the vertebrate arrestins, tree building was repeated with the optimized model settings for the nucleotide and amino acid input alignments given the additional constraint for *ARR0* to be monophyletic. For every model setting, several chains were combined after confirming that they converged to the same set of parameters with the help of Tracer v1.6 (Rambaut, Suchard, and Drummond, 2014) and logcombiner. Trees were analyzed with treeannotator and visualized in FigTree (Raumbaut, 2006).

3.2.5 Identification of specificity determining positions

For identification of specificity determining positions (SDPs) of closely related paralogs that arose from a recent duplication, respective sequences were grouped, aligned and filtered to contain a redundancy < 98 % and coverage > 70 %. The following groups were investigated: teleost *SAGa, b*, teleost *ARR3a, b*, teleost *ARRB2a, b*, all *ARR0* including sea urchin *ARR0.1*. The filtered alignments were analyzed with four complementary SDP detection tools, the entropy-based Sequence Harmony approach (SH, Pirovano, Feenstra, and Heringa (2006) and Feenstra et al. (2007)), the machine learning approach multiRELIEF (Ye et al., 2008; Brandt, Feenstra, and Heringa, 2010), Xdet, which is based on analysis of mutational behavior (Pazos, Rausell, and Valencia, 2006) and S3det based on multiple correspondence analysis (MCA, Rausell et al. (2010), section 1.4.3). The first two approaches were run via the webserver (Brandt, Feenstra, and Heringa, 2016), while the latter two are implemented in the program jdet 1.4.5. Positions retrieved with the default values of the respective programs (exception: S3det -m 2) were filtered according to the following, conservative cut-offs: SH z-scores $<$ -6, multi-RELIEF-scores $>$

0.7 and `Xdet-scores` < 0.6. Group distinction was computed automatically (unsupervised) in `S3det` except for teleost *ARRB2*. Positions were only considered as specificity determining if they were retrieved with at least two of the four methods.

3.2.6 Testing for natural selection

To test for natural selection, alignments of coding DNA sequences were constructed restricted to specific subbranches of interest (section 1.4.2). Regions encoding frame shift mutations, containing stop codons or gaps were excluded from further analysis. I excluded potential recombinant sequences by testing for recombination in the group alignments with the `RDP4` software (Martin et al. (2015), *SAGa, b* zebrafish (*Danio rerio*), *ARR3* stickleback). I assume that recombination and gene conversion can only occur within the same species and thus excluded incomplete lineage sorting for the species considered. Positive selection was tested on predefined foreground branches with the branch-site model of `codeml` inside the `PAML` program (Yang (2007), κ to be estimated, F3X4 and Codon table tested as Codon frequency models, see Fig. A.2 for exact trees and branches tested). The significance of difference of the maximum log-likelihoods of the null model ($H_0, \omega_2 = 1$) and the alternative model ($H_1, \omega_2 \geq 1$) was assessed by comparing the results of the Likelihood Ratio Test (LRT) with the χ^2 distribution of P-values (< 0.05). When the alternative model was significantly better than the null model, specific sites under positive selection were assessed according to the significance levels of the Bayes Empirical Bayes (BEB) method. Additionally, I performed bootstrapping and assessed the distribution and confidence intervals of the bootstrapped estimates with the `CODEML_SBA` method (Bielawski, Baker, and Mingrone (2016), Tab. B.5). Some data sets show a slightly bimodal distribution of ω_2 and/or p_1 and thus obtained rather uncertain parameter estimates (reported as μ, σ and upper and lower quartiles). The fraction of sites under positive selection (p_2) was calculated as follows: $p_2 = 1 - (p_0 + p_1)$.

3.2.7 Assessement of sequence conservation, conservation of posttranslational modification motifs and splice variant conservation

Sequence conservation was calculated with the Karlin score (Karlin and Brocchieri, 1996) implemented in `AACon` (Manning, Jefferson, and Barton, 2008) for alignments of individual orthology groups (*SAG*, *ARRB1*, *ARRB2*, *ARR3*) excluding lamprey sequences. To minimize the effect of missing data on conservation calculations, the alignments were filtered so that sequences with a coverage > 90 % remained.
Conservation of post-translational modifications (PTMs) were evaluated for those PTMs that were supported by more than one high-throughput phosphoproteomics data set (Hornbeck et al., 2015) or reported in a low-throughput arrestin-focused study. Conservation of kinase motifs was evaluated based on motifs extracted from the Human Protein Reference Database (Keshava Prasad et al., 2009).
Conservation of splice variants was assessed based on the isoforms annotated for the four arrestin paralogs in the `Ensembl genome browser` for human, cow or mouse (*Mus musculus*) or that were reported in literature (Tab. B.2). Conservation was evaluated based on the theoretical, genetic prerequisites (conservation of stop/start, reading frame, canonical splice sites, SS) for the respective isoforms and does not consider expression data from other species. Due to exon 16's short length (4 nt + stop codon in cow *SAG*), the p44 splice variant was not considered.

3.2.8 Parsimonious reconstruction of exon gain and loss events

Exon gain and loss events occurred several times at the same position within the arrestin gene family in deuterostomes. As a consequence, several scenarios exist with the same number of events (intron gains and losses). For reconstruction and mapping of exon loss and gain events, the number of events was minimized without resolving whether these are actually intron gains or losses considering the ongoing and unresolved discussion about introns-late vs. introns-early concepts (Rogozin et al., 2012). For counting the number of events, the root state was hypothesized to be the same as in fruit fly's phosrestin-1 and roundworm arrestin (*Caenorhabditis elegans*), which have no introns at the exon gain and loss hotspots, with exception of intron 138c in roundworm.

3.3 Results

3.3.1 Evolution of the arrestin fold family based on database inquiries

I aimed at first updating the inventory of proteins that harbor an *arrestin_N* and *arrestin_C* domain (PF00339, PF02752). For that purpose, I queried `UniProtKB` and `OrthoDB` animals in a `jackhmmer` search with pHMMs built from the four human arrestin full-length sequences. Three `jackhmmer` iterations maximized the number of human homologs in the sequence set, that contain an *arrestin_N* domain and *arrestin_C* domain. At the same time, inclusion of other domains was avoided, namley *DSCR3* and *VPS26B*, which harbor the Vacuolar protein sorting-associated protein 26 related domain (*Vps26*, PF03643) and are members of the arrestin clan as reported previously in Alvarez (2008) and Aubry and Klein (2013). The obtained full-length `jackhmmer` set was further filtered to exclude 142 sequences that did not have either of both domains (Fig. A.1). 86 % of the remaining sequences possess both an *arrestin_N* and an *arrestin_C* domain. Phylogenetic inference of sequences that are not homologous in full-length can potentially cause artifacts during phylogenetic inference. For this reason, an additional phylogenetic tree was generated based on the single `Pfam` arrestin domain models (PF00339, PF02752, Fig. A.3). Both trees have a very similar topology, although differences exist. The query with single domain `Pfam` models retrieves a higher number of hits that cover more species in comparison to the full-length models (Tab. 3.1), although the same clades on the phylogeny are covered. As the single domains are not found in more ancestral clades than the linkage of both domains, I conclude that this is likely an effect of missing data and incomplete annotation of `UniProtKB` entries in regard to sequence coverage rather than reflecting the loss of the linkage of the *arrestin_N* and an *arrestin_C* domains. I thus confirm the linkage of both domains throughout the phylogeny, which might be reminiscent of an early duplication of a single ancestral domain.

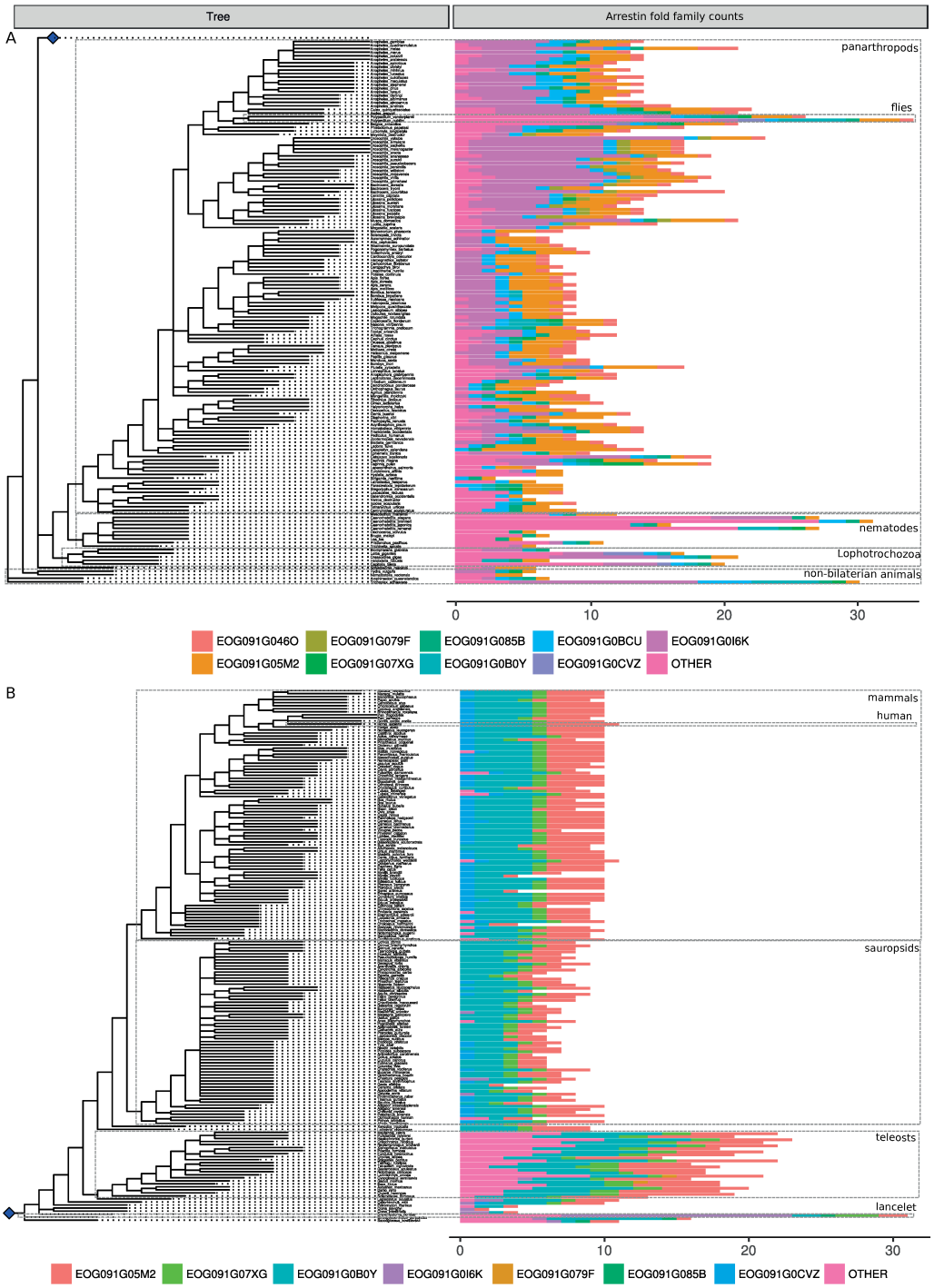

Figure 3.1: Abundance of arrestin fold family members in animals and orthology assignment according to OrthoDB. Hits were mapped to the NCBI taxonomy of animals. Deuterostomes are represented by a blue diamond in A and extensively shown in B. The color coding corresponds to different orthology groups. Note that groups with < 29 members were collapsed to the single group "Other". Clades of interest are pointed out on the right side.

Table 3.1: Scan of the `UniProtKB` database with arrestin profile Hidden Markov Models (pHMMs). The respective hits were filtered as described in section 3.2.1. The number in parenthesis refers to the number of unique sequence IDs that retrieved at least one domain hit. Note that hits that were not assigned to a specific species in `UniProtKB` were removed during the filtering process.

Database name	pHMM	# Hits retrieved	# Species covered	# Species covered outside of animals	Orthology groups
`OrthoDB`	`jackhmmer` full-length	3487	330	0	109
`UniProtKB`	`jackhmmer` full-length	2389	357	63	NA
`UniProtKB`	*arrestin_N* domain	3190 (3150)	625	42	NA
`UniProtKB`	*arrestin_C* domain	3416 (3395)	629	30	NA

The obtained set of homologs encompasses ten members in human for both the `UniProtKB` and `OrthoDB`, in accordance with Alvarez (2008) (*ARRDC1-5*, *TXNIP*, *SAG*, *ARRB1*, *ARRB2*, *ARR3*) and will be referred to as the arrestin fold family in the following. In the following, I first describe the statistics based on the scan of animal `OrthoDB` (Fig. 3.1), which is more complete in respect to paralog counts than `UniProtKB` with an average count of 9.8 and 7.2 arrestin fold family members per species, respectively. Please see Fig. A.4 for arrestin paralog counts in bilaterians based on `UniProtKB`. Second, I evaluate the abundance of arrestin fold family members outside of animals based on scanning `UniProtKB` with the full-length models (Fig. 3.2). I return to the differences of both databases and annotations derived with the EMS-pipeline in the discussion (section 3.4.1).
The arrestin fold family is part of the *Arrestin N-like clan* (CL0135) as defined by `Pfam 31.0`, which corresponds to the arrestin clan described in the literature. The *Arrestin N-like clan* includes the following domains: *arrestin_C*, *arrestin_N*, *Spo0M*, *Vps26*. It exceeds the literature classification by inclusion of the domains *LDB19* and *Bul1_N* (both restricted to fungi). The arrestin fold family has 3487 members that belong to 109 different orthology groups on the highest clade level available in `OrthoDB` (animals, Tab. 3.1, Fig. 3.1). Nine orthology groups have at least 29 members and cover 88 % of all animal arrestin fold family members. The majority of vertebrate arrestin fold proteins belongs to one of the following four `OrthoDB` orthology groups, that contain the following human genes: *SAG/ARRB1/ARRB2/ARR3* (arrestins, EOG091G05M2), *ARRDC1* (EOG091G07XG), *ARRDC2/ARRDC3/ARRDC4/TXNIP* (EOG091G0B0Y) and *ARRDC5* (EOG091G0CVZ).
The monophyly of those orthology groups is also supported by phylogenetic inference with both full-length and single domain sequences as extracted from the `UniProtKB` database (Fig. 3.3) and the exon–intron structure of individual paralogs from human. Arrestins as well as the *ARRDC2-4/TXNIP* group strictly conserve the exon–intron structure within the respective groups (arrestins: 13-16 exons, *ARRDC2-4/TXNIP*: 8 exons). *ARRDC1* shares three exon–intron boundaries with the *ARRDC2-4/TXNIP* group supporting *ARRDC1* as the closest outgroup to *ARRDC2-4/TXNIP* as inferred from the phylogenetic inference. *ARRDC5* shares the two existing exon–intron boundaries with both *ARRDC1* and *ARRDC2-4/TXNIP* groups. The origin

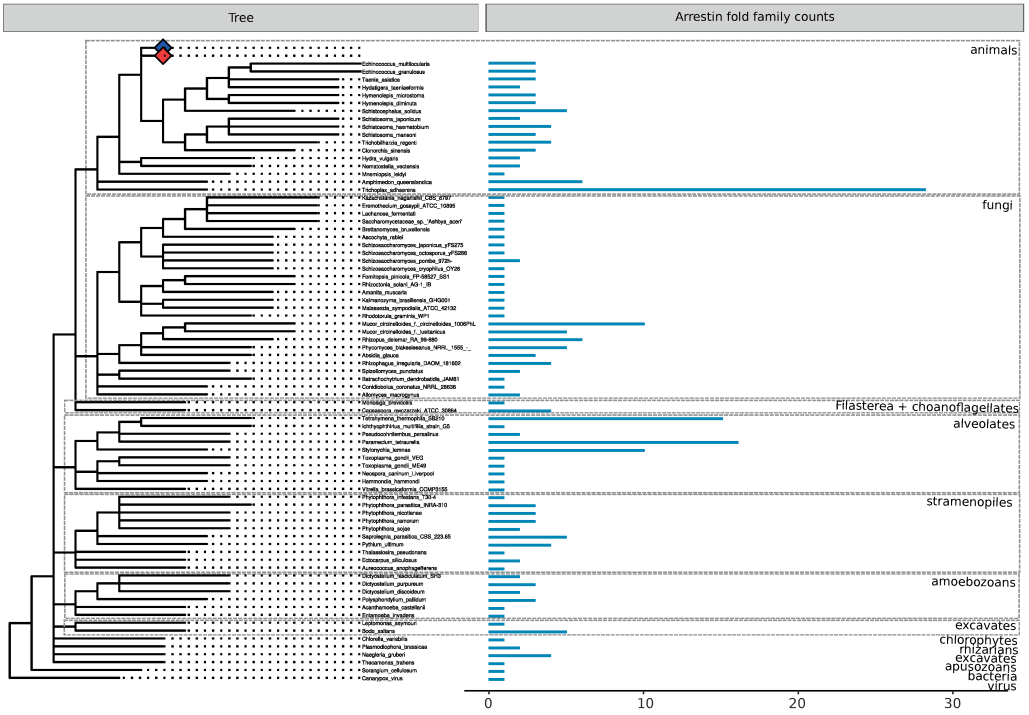

Figure 3.2: Abundance of arrestin fold family members in different domains of life according to `UniProtKB`. Hits were assigned to the arrestin fold family if they contained at least one *arrestin_N* or *arrestin_C* domain (section 3.2.1) and were mapped to the `NCBI` taxonomy. The counts for *Phytophthora sojae*, but not for the strain P6497 of the same species are shown, although both species were included for phylogenetic inference. The blue and red diamond simplify the groups of protostomes and deuterostomes, respectively (see Fig. A.4 for details). Clades of interest are pointed out on the right side.

of the amniote-specific *ARRDC5* group waits to be elucidated given its placement with about equal distances to arrestins and *ARRDC1-4/TXNIP* in the phylogenetic tree (Fig. 3.3). Three lophotrochozoan sequences that are part of the same orthology group (EOG091G0CVZ) were identified to be false group assignments provided by `OrthoDB` after manual inspection (Fig. 3.1). The putative *ARRDC5* paralogs in fly and worm as proposed by Aubry and Klein (2013) do not represent 1:1 orthologs to amniote *ARRDC5*. Two of the three proposed orthologs were excluded from the `UniProtKB` set due to filtering, while the third sequence clusters with another clade. In `OrthoDB`, all three sequences are assigned to other, small orthology groups. This study is the first to report that *ARRDC5* is specific to amniotes, which is further supported by the phylogenetic inference and orthology information extracted from the `Ensembl Compara database` (Vilella et al. (2009), `Ensembl Version 89`). Ray-finned fish are a sister-class of lobe-finned fish that together form the bony fish, a major clade within vertebrates. The majority of living representatives of ray-finned fish fall into the infraclass of teleosts with only the nuclear genome of spotted gar (*Lepisosteus oculatus*) sequenced outside teleosts. As expected due to the teleost-specific 3R-whole genome duplication (WGD), teleosts have about twice as many orthologs as lobe-finned fish or spotted gar (Fig. 3.1, A.4). This number is further increased by at least one additional ray-finned fish-specific paralog (classified as "Other") that forms a monophyletic group outside of the *ARRDC2-4/TXNIP* group. Surprisingly, its *arrestin_N* domain is clustered more distantly in respect to the *ARRDC2-4/TXNIP* group than the *arrestin_C* domain of the same sequence in the single domain trees,

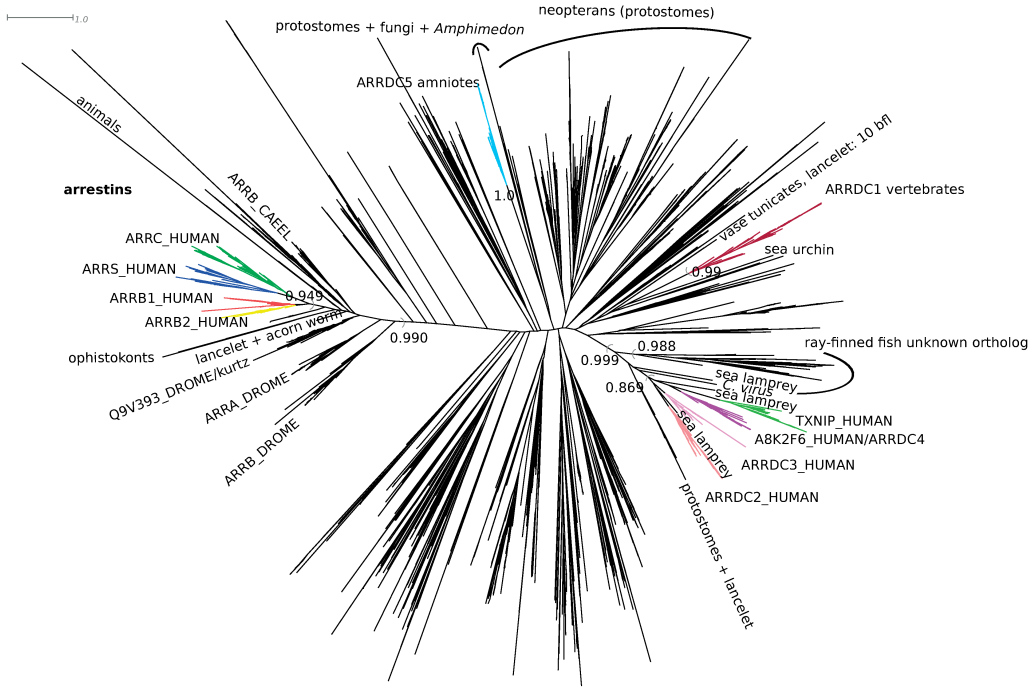

Figure 3.3: Approximate maximum likelihood tree of the full-length arrestin fold family members as extracted from `UniProtKB`. Hits were assigned to the arrestin fold family if they contained at least one *arrestin_N* or *arrestin_C* domain (section 3.2.1, Tab. 3.1). The tree was generated with the `FastTree` software and bootstrapping was performed 1,000 times with `SeqBoot` (Felsenstein, 2017). Bootstrap values are shown for splits that contain groups with vertebrate arrestin fold members. The groups with human arrestin fold protein members and their respective 1:1 orthologs are marked in color. Vertebrate arrestins clearly form a monophyletic group within the arrestin fold family. The well known arrestins from worm and and fruit fly are labeled with their `UniProtKB` IDs. All deuterostome representatives as well as other subtrees of interest are labeled.

possibly indicating faster evolution of the *arrestin_N* domain (Fig. A.3). The arrestin and *ARRDC2-4/TXNIP* groups expanded in the vertebrate ancestor with generally lower paralog numbers outside of vertebrates. A striking lineage-specific extension occurred in lancelet (*Branchiostoma floridae*), that possesses the highest count of arrestin fold proteins in deuterostomes with 31 homologs according to `OrthoDB` (Fig. 3.1). Most members belong to a group that clusters outside of vertebrate *ARRDC1*.
The clade with most representatives within animals are Bilateria (bilaterians), which is composed of the sister-superphyla deuterostomes and protostomes. The encoded gene copy number of arrestins greatly varies in protostomes with numerous lineage- or clade-specific extensions as seen by the appearance of separate orthology groups ("Other") and clade-specific subtrees in the phylogenetic tree (e. g. neopteran-specific, Fig. 3.1, 3.3). The lineage-specific extensions in nematodes (Caenorhabditis) and flies (Polypedilum) result in an increase in up to 30 arrestin fold family homologs as described by Alvarez (2008) and Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014). Subtrees for both clades with most members are located in proximity, but not within the *ARRDC5* subtree. The emergence of three groups, arrestins, *ARRDC1* and *ARRDC2-4/TXNIP*, predates the emergence of bilaterians according to `OrthoDB`. At least two of the four surveyed non-bilaterian animals also possess members of the EOG091G0I6K, EOG091G0BCU and EOG091G085B groups that do not have any representatives in human.

To determine the existence of arrestin homologs in even earlier branching species outside of animals, I considered the results of the scan of full-length `jackhmmer` pHMMs against `UniProtKB`, which covers more species than `OrthoDB`, but is lacking the assignment to orthology groups (Fig. 3.2, Tab. 3.1). The hits against the `UniProtKB` database with the full-length query cover the clades of animals, fungi, amoebozoans, alveolates, excavates and stramenopiles with at least three species representatives of each of these clades. I additionally detected hits in the following clades with one representative each: bacteria (*Sorangium cellulosum*), virus (*Canarypox virus*), chlorophytes (*Chlorella variabilis*), rhizarians (*Plasmodiophora brassicae*) and apusozoans (*Thecamonas trahens*). This study confirms the absence of arrestin fold proteins in embryophytes and their low abundance in chlorophytes described by Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014). In comparison to Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014), I miss arrestin fold family members in glaucophytes and haptophytes, although, the haptophytes *Emiliania huxleyi* is recovered in the single domain scan with the *arrestin_N* domain. My phylogenetic inference also confirms that the arrestin fold protein in *Canarypox virus* probably originated from horizontal gene transfer of a vertebrate member of the *ARRDC2-4/TXNIP* group (Fig. 3.3, Aubry and Klein (2013)). Arrestins clearly form a monophyletic group within the group of arrestin fold proteins, which expanded in deuterostomes to give rise to the four paralogs seen in human. Arrestins can be traced back to the holozoan orders of choanoflagellates (*Monosiga brevicollis*) and Filasterea (*Capsaspora brevicollis*) outside of animals in agreement with Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014), while the *ARRDC1-5/TXNIP* group is limited to animals. All other arrestin fold proteins outside of opisthokont cluster into groups that are about equally distant from arrestins and the *ARRDC1-5/TXNIP* group and do not have 1:1 orthologs in human.

3.3.2 The `ExonMatchSolver` annotation of arrestins is more complete than arrestin database entries in `OrthoDB`

I applied the EMS-pipeline to improve the annotation of arrestins in deuterostomes. Arrestins possess 13-16 exons. Their exon–intron structure is conserved even across the four paralogs with minor deviations, which is an important prerequisite for application of the EMS-pipeline (Chapter 2). Furthermore, arrestins are well characterized within mammals (section 1.5.4) simplifying the creation of a high quality alignment, which is used for building the initial pHMMs (section 3.2.4).
In fact, I demonstrate that the application of the EMS-pipeline is a more successful strategy to trace the details of arrestin evolution in comparison to a coarse database analysis. I compare the detailed arrestin annotations in deuterostomes, which are assumed to be correct as assessed by expert inspection, with the counts of the `OrthoDB` group EOG091G05M2 (arrestins). Although `OrthoDB` is more complete than `UniProtKB`, the arrestin paralog number deviates in 44 % of all cases from the curated annotations. Specifically, `OrthoDB` under- and overpredicted the number of paralogs in 20 and five of 57 species, respectively. In general, the EMS-pipeline found paralogs that are missing from `OrthoDB` (Fig. 3.4). `OrthoDB` overpredicted sequences due to mis-assembly (in pig, *Sus scrofa*), inclusion of a pseudogene (in opossum, *Monodelphis domestica*), a naming mistake (in human), and included two additional sequences without any further reference (in lancelet and acorn worm, *Saccoglossus kowalevskii*). The curated annotation is in general more complete than the respective database entries regarding sequence coverage as a result of consideration of fragmented gene loci. Furthermore, it benefits from a fundamental improvement of the annotation of SSs, short and terminal exons.

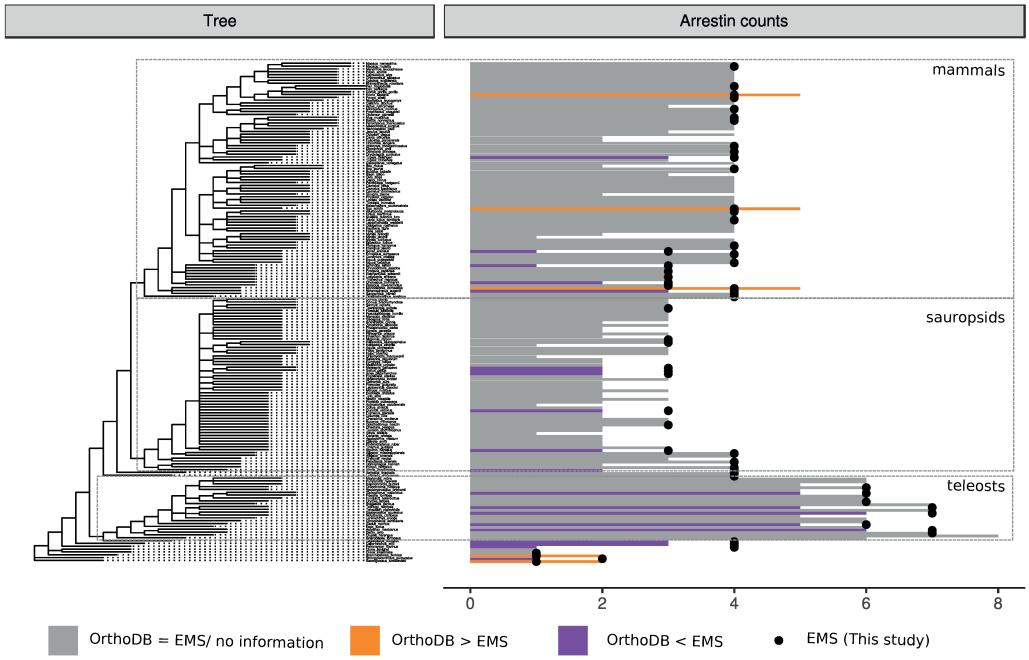

Figure 3.4: Comparison of number of arrestin paralogs in deuterostomes between the `ExonMatchSolver` (EMS)-pipeline and `OrthoDB`. Higher and lower paralog counts were obtained for the `OrthoDB` group EOG091G05M2 (arrestins) by genome mining in combination with manual curation for 20 species (purple) and five species (orange), respectively, as compared to the `OrthoDB`. The paralog counts and annotations obtained with the EMS-pipeline and that are based on an expert opinion are assumed to be correct. `OrthoDB` overpredicted sequences due to mis-assembly (pig, *Sus scrofa*), inclusion of a pseudogene (opossum, *Monodelphis domestica*), a naming mistake (*Homo sapiens*), included two additional sequences without any further reference (*Branchiostoma floridae*, *Saccoglossus kowalevskii*). Five additional genomes were mined with the EMS-pipeline that are not included in the `OrthoDB` (not shown).

Furthermore, I added five species critical to resolve the arrestin genealogy that were not included in `OrthoDB` (green sea urchin, *Lytechinus variegatus*; bat star, *Patiria miniata*; little skate, *Leucoraja erinacea*; arctic lamprey, *Lethenteron camtschaticum* and aardvark, *Orycteropus afer afer*). The inclusion of those species permits a more precise mapping of the following duplication and loss events on the species tree: the cartilaginous fish-specific duplication of *SAG*, the duplication of *ARR0* in the ancestor of sea urchins and the loss of *ARR3* in the ancestor of afrotherians/xenarthrans. Furthermore, the inclusion of the green sea urchin empowered me to trace positive selection and to identify SDPs for the respective branches and duplicated arrestins. This Chapter thus demonstrates how detailed curation can change and improve the detailed duplication and deletion history of an individual gene.

3.3.3 Arrestin paralog gain and loss patterns based on the `ExonMatchSolver` annotations

The vertebrate 2R-WGD leads to the emergence of four arrestin paralogs

The arrestin sequences retrieved from the deuterostome genomes excluding lamprey sequences fall into five well separated and supported orthology groups with > 85 % bootstrap support values (BS)/ posterior probability (pp) in all ML and Bayesian phylogenetic trees (Fig. 3.5, 3.6, A.5). Please consider the next subsection for details

about lamprey arrestins. Four of the five groups contain one of the four human arrestins each. The fifth group, *ARR0*, is formed by non-vertebrate arrestins with intermediate properties between the visual and non-visual types and encompasses the previously cloned lancelet arrestin. *ARR0* is most similar to the non-visual vertebrate arrestins, especially to *ARRB1* (average identity of all *ARR0* to human *ARRB1* 61.9 %). Each of the four gene trees of the vertebrate orthology groups is in good accordance with the vertebrate species tree. Especially the *SAG* subtree (dark blue in Fig. 3.5, 3.6) perfectly resolves the vertebrate species phylogeny except for few splits. The subtrees of the non-visual types lack this high resolution due to its lower substitution rates in comparison to visual arrestins. Nevertheless, important clades like ray-finned fish and lobe-finned fish including birds, mammals and amphibians are resolved (Fig. 3.5, 3.6, A.5).
The arrestin gene trees furthermore show that the visual arrestins, *SAG* and *ARR3*, form a well supported monophyletic group with 100 % BS and 100 pp. The monophyly of non-visual arrestins, *ARRB1* and *ARRB2*, is less well supported (23.5 % BS and at least 88.9 pp). Branch lengths to the shared ancestor are short within the ML tree. In order to check that this tree topology is not the result of convergent evolution of visual arrestins, I removed the alignment columns that are known to mediate receptor binding (Vishnivetskiy et al. (2004), Hanson et al. (2006), Zhan et al. (2011a), Vishnivetskiy et al. (2011), Szczepek et al. (2014), Ostermaier et al. (2014), and Kang et al. (2015), Tab. 1.3). The truncated alignment still produces the same tree topology in respect to the major splits mentioned above (Fig. A.6). The presented data thus supports the existence of one visual and one non-visual proto-arrestin derived from a single arrestin, *ARR0*. *ARR0* subsequently gave rise to two arrestins each (Fig. 3.7 B). Surprisingly, removal of the receptor specificity columns led to the resolution of the clades of lobe-finned and ray-finned fish in the *ARR3* subtree, that are not monophyletic in the full-length tree. This might hint at convergence in evolution of receptor specificity binding residues of this visual arrestin in mammals and ray-finned fish (Fig. A.6). Further research is required to resolve this issue.

Arrestin inventory in lampreys

In order to pinpoint the exact timing of the divergence of the four vertebrate arrestins, I focused on arrestins in available genomes of sea lamprey (*Petromyzon marinus*) and arctic lamprey. Jawless fish (cyclostomes), including lampreys, are the sister clade of the jawed vertebrates (gnathostomes). Lampreys experienced a poorly understood process of programmed DNA loss in their somatic cells corresponding to about 20 % of the germline DNA including protein-coding DNA (Smith et al., 2012). In order to take this process into account, I investigated germline and somatic genomes for sea lamprey, the only lamprey species for which both resources are available, and the germline genome of arctic lamprey. As expected, the number of paralogs retrieved from the germline and somatic genome of sea lamprey differ, encoding four and one arrestin paralog, respectively. While two non-visual arrestins were annotated without difficulty in the arctic lamprey, annotation of visual arrestins in the same species turned out to be problematic. The putative locus of *ARR3* was extremely fragmented with 12 exons situated on six different contigs. Nevertheless, predictions were consistent with the results of Kawano-Yamashita et al. (2011), who cloned one non-visual arrestin and one visual arrestin from arctic lamprey's pineal organ. The phylogenetic inference reveals that those two non-visual and one visual arrestin are 1:1 orthologs to the sea lamprey arrestins. The sea lamprey germline genome encodes an additional species-specific non-visual paralog. This arrestin is

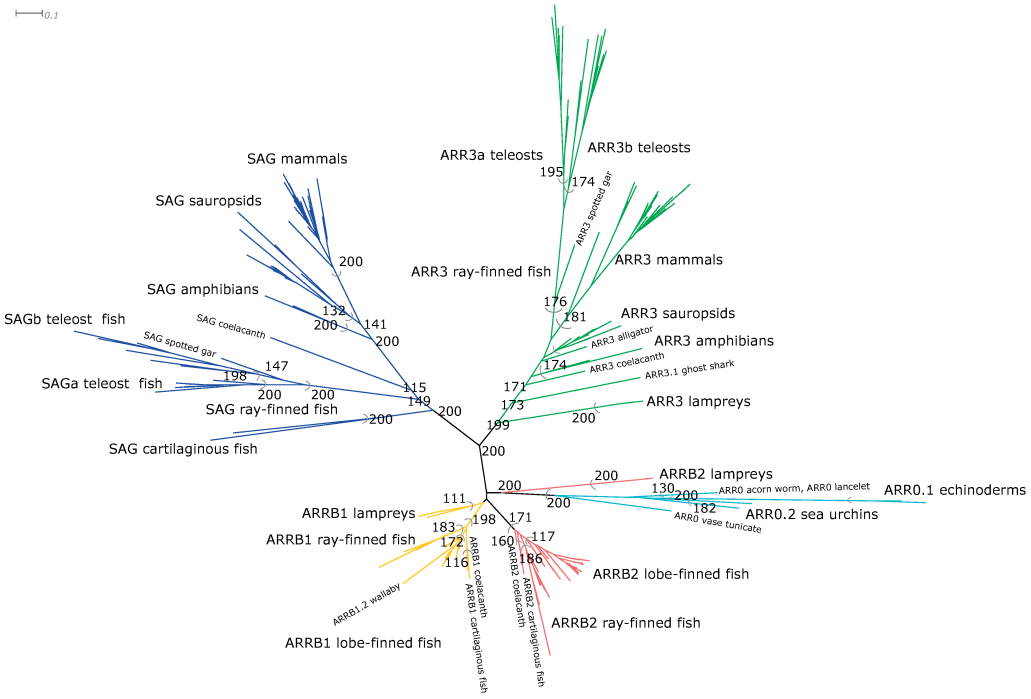

Figure 3.5: Unrooted maximum likelihood tree of arrestins. The tree was constructed from an amino acid alignment of deuterostome arrestins using `PhyML` (model JTT+I+G with α 1.04, 5 % of invariable sites and 200 bootstraps). The different monophyletic and well supported orthology groups are highlighted in different colors. Bootstrap support values from 50...100 % are shown for the labeled monophyletic groups. The phylogenetic tree was visualized with `Dendroscope 3.5.7` (Huson and Scornavacca, 2012).

most similar to *ARRB1*. Apart from those paralogs, I detected some more exons in the germline genomes of both lamprey species that are most similar to non-visual arrestins. However, exons orthologous to *SAG* remain missing in either lamprey genome. As this study cannot distinguish between an actual loss and missing data, the arrestin inventory for lampreys remains incomplete.

I included all those lamprey arrestins with a high quality gene annotation in the phylogenetic inference (Fig. 3.7, Tab. 3.2). Nevertheless, the position of the lamprey arrestins in the arrestin gene tree is difficult to resolve. While the visual arrestin from arctic lamprey clusters together with jawed vertebrate *ARR3* with high support in all trees, the position of lamprey non-visual arrestins varies depending on the tree building method applied (Tab. 3.2). Within the Bayesian nucleotide tree, both non-visual lamprey arrestins form a well supported monophyletic group (99.9 pp) that clusters with the jawed vertebrate *ARRB2*. In contrast, lamprey *ARRB2* clusters with *ARR0* with high BS (100 %) in the amino acid ML tree. The splits of the lamprey non-visual arrestins with their putative vertebrate orthology groups have weak support considering the other trees. The latter topology and the position of *ARR3* within the tree is in agreement with a 2R-WGD that is shared between jawed and jawless vertebrates (Fig. 3.7 B, Scenario 1). The other two tree topologies explained above rather support a shared first WGD and an independent lamprey-specific duplication (Fig. 3.7 B, Scenario 2).

The exact timing of the emergence of the four arrestin paralogs and thus the exact timing of the first and second round of the 2R-WGD cannot be resolved unambiguously with the available data. I return to this issue in the discussion (section 3.4.2).

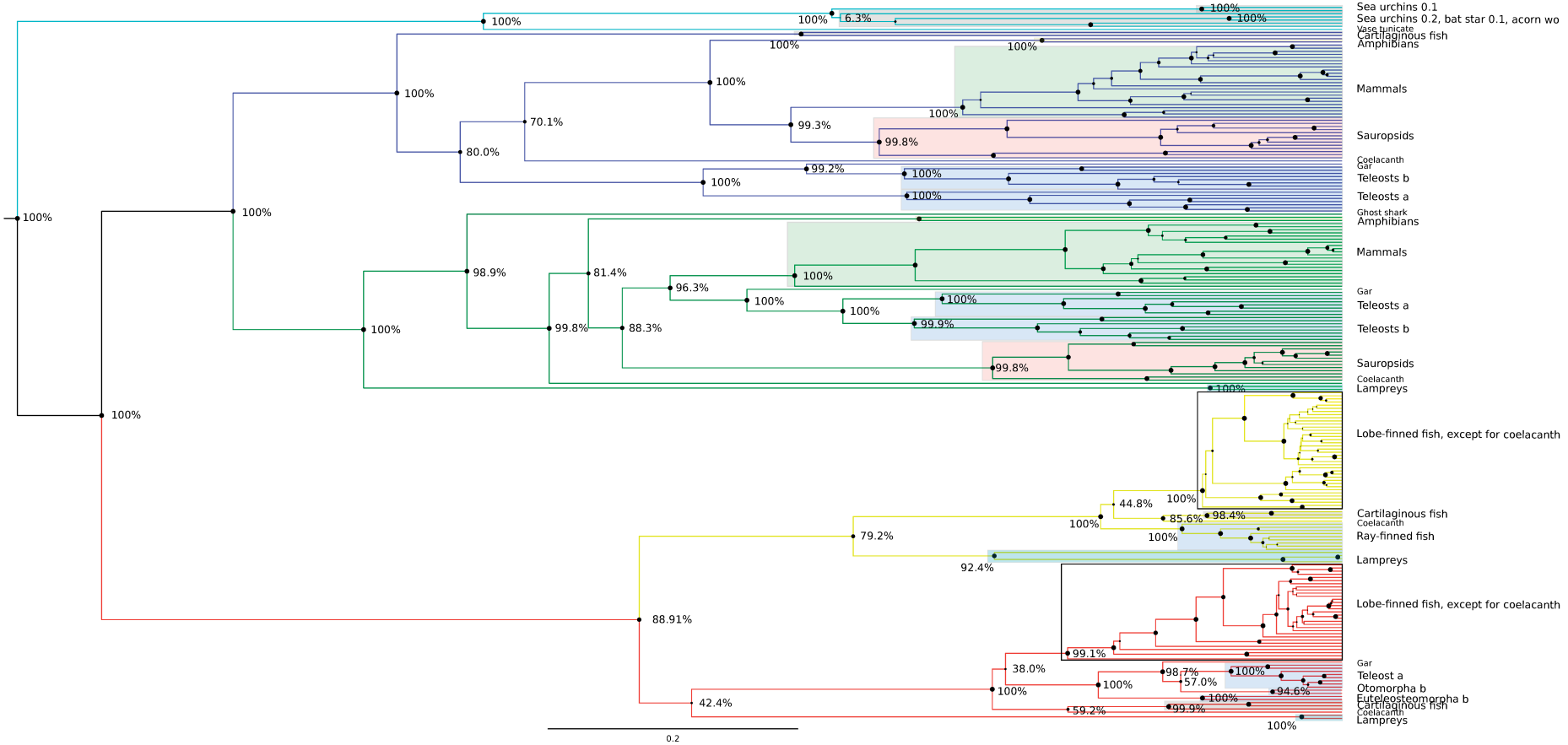

Figure 3.6: Rooted bayesian tree of arrestins. The tree was constructed from an amino acid alignment of deuterostome arrestins using `BEAST2` (birth–death model with relaxed log normal molecular clock, substitution model: JTT+I+G with α 1.04, 5 % of invariable sites model), birth–death rate parameters $\alpha = 1$ and $\beta = 10$). Furthermore, vertebrate arrestins were constrained to be monophyletic. The branches of the five main orthology groups are drawn in different colors: *ARR0* – turquoise, *SAG* – blue, *ARRB1* – yellow, *ARRB2* – red, *ARR3* – green. All highlighted species groups are monophyletic except for some *ARR0*, which are summarized for better readability. The size of the dot at all nodes scales to the respective posterior probability.

Table 3.2: Position of lamprey arrestins in phylogenetic inferences. The table provides the name of the orthology group and the bootstrap support values or posterior probabilities for inclusion of the respective lamprey paralog within this group. The putative non-visual lamprey arrestins (*ARRB1/2*) have different positions in the phylogenetic trees depending on the method applied. Abbreviations: ML – Maximum likelihood; B – Bayesian inference; AA – amino acid; jv – jawed vertebrate; NT – nucleotide; pp - posterior probability.

Tree type	***ARR3* lampreys**	***ARRB1* lampreys**	***ARRB2* lampreys**
ML AA	jv *ARR3*, 99.5 %	jv *ARRB1*, 40 %	*ARR0*, 100 %
ML AA without receptor specificity	jv *ARR3*, 96 %	*ARRB1/ARR0*, 36.5 %/29 %	jv *ARRB2*, 36.5 %
B AA	jv *ARR3*, 100 pp	jv *ARRB1*, 79.2 pp	jv *ARRB2*, 42.4 pp
B NT	jv *ARR3*, 99.9 pp	Monophyletic group (99.9 pp) within the jv *ARRB2* subgroup (48.6 pp)	

Tandem duplication of *ARR0* in sea urchins

The genomes of the majority of investigated non-vertebrate deuterostomes encode a single *ARR0* gene (Fig. 3.7 A). A notable exception are three echinoderm genomes. The purple and green sea urchins *Strongylocentrotus purpuratus* and *Lytechinus variegatus* possess two paralogous *ARR0* genes, which are located about 110 kb apart from each other and have a mean sequence identity of 61 %. This arrangement is indicative of a tandem duplication. The sea urchin *ARR0.1* genes show an accelerated substitution rate in comparison to *ARR0.2* and bat star *ARR0.1* as indicated by long branch lengths within the ML tree. They are also identified as a separate group in unsupervised clustering (section 3.2.5). The ancestral echinoderm *ARR0* had already diverged to some extant from the other *ARR0* before the gene got duplicated in the ancestor of sea urchins as indicated by the position of bat star *ARR0.1* in the phylogenetic inference and its clustering separately from the two main groups in sequence space. Sea urchin *ARR0.1*s carry SDPs that are distinct from homologous positions in all other investigated *ARR0*s (Fig. 3.8). Interestingly, the amino acid in the *ARR0* main group is often identical to the amino acid at the homologous position in cow *ARRB1*. Some of the sea urchin substitutions lead to the charge reversal of phosphate sensing (R165E, Vishnivetskiy et al. (2011)), inositol-hexa-phosphate (IP6) binding (K157V, Milano et al. (2006)) and adapter protein-2 (AP-2) binding (Burtey et al., 2007) residues (R395C, Fig. 3.8 A, C, D). Furthermore, receptor binding residues differ (Fig. 3.8 B). After the tandem duplication and before speciation of both sea urchin species, different fractions of sites evolved under positive selection in *ARR0.1* and *ARR0.2*, 15 % and 5 %, respectively (Fig. A.2 A, Tab. B.6, B.5). This reflects the asymmetrical evolution of both paralogs after the duplication and hints at neofunctionalization of *ARR0.1*. Among those positively selected residues are positions involved in or neighboring to receptor binding sites as well as to IP6 binding residues in the *ARR0.1* branch, which were also identified as SDP (Tab. 3.3). Further inspection of the sequence space revealed that vase tunicate (*Ciona intestinalis*) *ARR0* might be functionally different from the main group. This notion is also supported by the *ARR0* gene tree, which is not in accordance with the species tree and where vase tunicate *ARR0* clusters outside of the main group (Fig. 3.5, 3.6, A.5, A.6).
Furthermore, I find two largely identical *ARR0* sequences in the bat star genome (exonic nucleotide sequences are 98.7 % identical, intronic nucleotide sequences are

89 % identical). Clearly, these two copies are the result of a very recent duplication independent of the duplication event that generated the much older paralogs in the sea urchins.

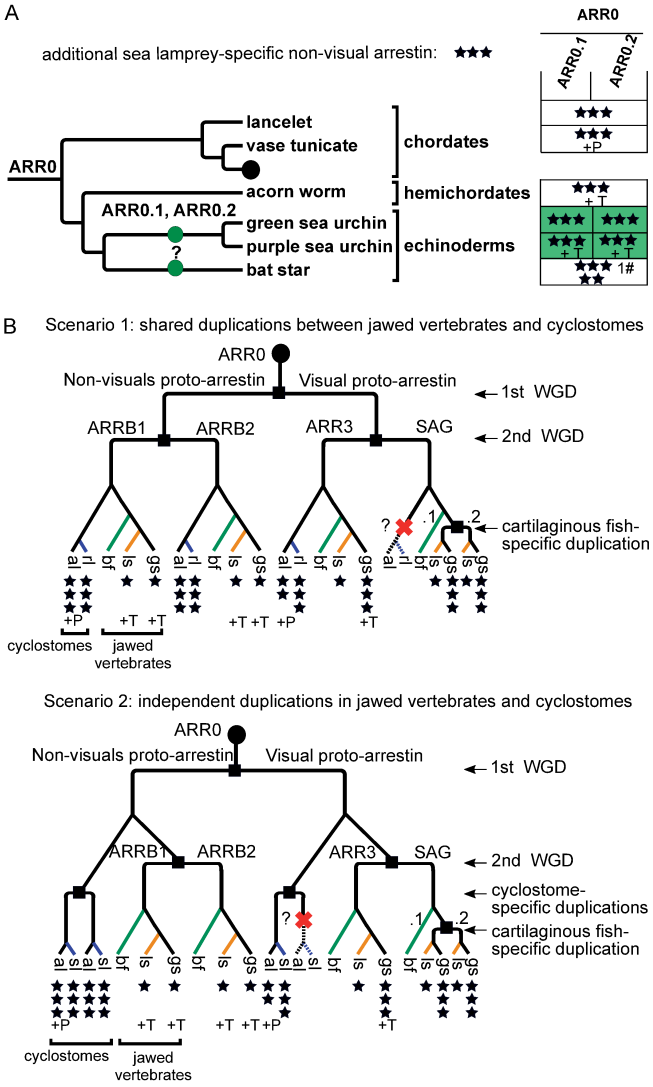

Figure 3.7: Duplication and deletion of arrestin paralogs in early-branching deuterostomes. A – Species tree of early-branching deuterostomes with mapped duplication events of arrestins (dots). B – Schematic arrestin gene tree for vertebrates (square in A). A cross indicates a gene loss. New gene names are given above the dot or branch. The gene loss/ duplication pattern was simplified for bony fish (bf), see Fig. 3.10, 3.13 and A.7. The completeness of arrestin annotations in the respective genomes is depicted with three stars indicating zero to three missing exons, two stars four to eight missing exons, one star more than eight missing exons and dash (-) that no gene fragments were detected. Additional support from other omics-data for cartilaginous fishes and jawless fishes and from experimentally validated `Genbank` entries is indicated by the following abbreviations: T - transcriptome evidence; P - protein evidence. The hash (#) indicates the number of frame shift mutations contained in the exon annotation. Note that the order of jawless fish-specific and cartilaginous fish-specific duplications in relation to each other was chosen arbitrarily. The additional non-visual arrestin detected in the germline genome of sea lamprey was not included in the scenario. Phylogenetic trees were created with `Treegraph 2.0.54` (Stover and Muller, 2010).

Table 3.3: Positively selected residues of arrestins detected with the Bayes Empirical Bayes (BEB) method. The branch-site model of the PAML package was used to identify sites under positive selection in the specified foreground branch. The position in column two refers to the position within the group alignment, while the homologous position in cow serves as a reference. The position in *ARR0* is given in respect to *ARRB1* in cow. The function assignment is based on literature review. See Tab. B.8 for further details. Positions that were also identified as specificity determining position (SDP), are marked by a cross. SDPs were not determined for all subgroups as indicated by "NA".

Foreground branch	Pos. in cow ortholog	Function known from homologs	SDP?
ARR0.1 sea urchins	N83	second neighboring to receptor binding residue	x
	E102	-	x
	K157	low affinity IP6 binding site	x
	N162	neighboring to low affinity IP6 binding site	x
	N225	second neighboring to receptor binding residue	-
	C242	receptor binding	x
	N382	second neighboring to clathrin binding site	-
ARR0.2 sea urchins	P89	neighboring to PxxP motif	-
SAG.1 ghost shark	K2	-	NA
	P134	neighboring to receptor binding residue	NA
	R171	phosphate sensor	NA
	G185	neighboring to PxxP motif	NA
	G217	-	NA
	E262	receptor binding	NA
	N305	second neighboring to polar core	NA
	T334	second neighboring to high affinity IP6 binding site	NA
	G27	second neighboring to receptor binding residue	NA
SAGb teleost	V35	second neighboring to polar core	x
	W194	receptor binding	-
SAGb Acanthopterygii	P93	neighboring to PxxP motif	NA
	A180	neighboring to PxxP motif	NA
	S210	-	NA
ARR3b euteleosts	M55	neighboring to μ2 adaptin binding site	x
	F254	neighboring to receptor binding residue	x

Tandem duplication of *SAG* in cartilaginous fishes

The clades of bony fish (comprising reptiles, birds, fish and mammals) and cartilaginous fish including the chimaeras, sharks, and rays together form the jawed vertebrates. I identified two copies of *SAG* in ghost shark, the only available chimaera

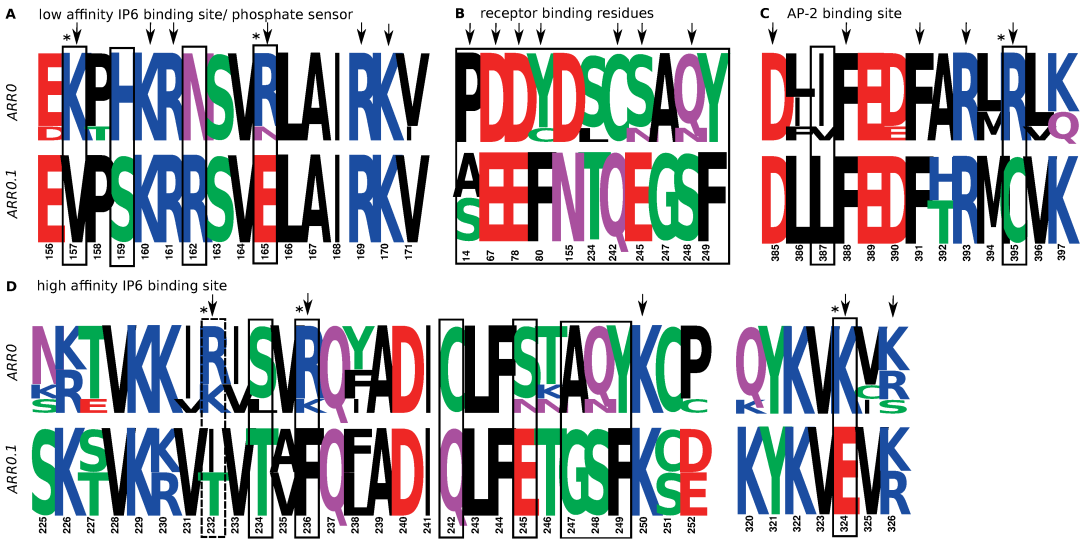

Figure 3.8: Specificity determining positions discriminating between sea urchin *ARR0.1* and all other considered *ARR0*s including sea urchin *ARR0.2*. Amino acid frequency logos are shown for *ARR0* and *ARR0.1* of sea urchins ordered by functionality of motifs known from studies in vertebrate arrestins (A-D). Positions that are known to directly confer the respective functionality are marked by arrows. Some mutations change the charge of the respective residue (marked by *). Positions identified by specificity determining position (SDP) analysis are highlighted by black boxes. As receptor specificity is mediated by a rather big interface, only the SDPs are shown that are known to be involved in receptor binding and their direct neighbors. An additional position that shows differences in both groups (manually identified) and is associated with the respective function is highlighted by a dotted box. The numbering of the positions refers to bovine *ARRB1*. Please see Tab. 1.3, B.8 for literature references of functions. The figure was created with `Weblogo` (Crooks et al., 2004).

genome. The two copies, *SAG.1* and *SAG.2*, are arranged in tandem about 8 kb apart on opposite strands. With the help of the EMS-pipeline and additional manual curation, I confirmed that a second *SAG* gene is also encoded in the draft assembly of the genome of the little skate. Therefore, the tandem duplication of *SAG* predates the split of chimaeras and sharks between 413-473 mya (Fig. 3.7 B, A.8). The annotation of arrestins in the little skate genome turned out to be especially problematic, as the genome was highly fragmented with a contig N_{50} of 665 nt. Only fragments of arrestins (1-4 exons) were found to be situated on the same genomic fragment. Here, I again took advantage of the prior information about the conserved exon–intron structure of vertebrate arrestins and estimated the number of paralogs based on different hits of one exon-family of sufficient length (exon 5). In little skate, five complete and one partial exon 5 were detected with *E*-values $<= 1.4e-05$, whereby the incomplete exon 5 was located at the end of contig *AESE011046971.1* (Fig. 3.9). The observation of at least five paralogs is further supported by the detection of five reliable sequences for exons 3 and 9 each (*E*-values $<= 4.4e-07$ and $<= 1.9e-04$, respectively). I confirmed that this additional paralog is in fact a second *SAG* by constructing a ML tree of the nucleotide sequences of exon 5 from little skate, spotted gar and ghost shark. As expected, one exon 5 sequence from little skate clustered together with *SAG.1* and *SAG.2* from ghost shark, respectively, forming a monophyletic group with *SAG* from spotted gar (Fig. 3.9).

The protein sequences of arrestin-1.1 and arrestin-1.2 of ghost shark have about equal

fractions of identical amino acids with 51 % and 55 %, respectively, to the single arrestin-1 of spotted gar. This is further reflected by about equal substitution rates of *SAG.1* and *SAG.2* since their duplication (Fig. 3.5). Positive selection has been acting on about 13 % of sites in ghost shark *SAG.1* (Tab. B.5, Fig. A.2 B). The specific sites identified to be under positive selection are conserved among all other *SAG* and even other arrestin paralogs. Among those are two residues involved or directly neighboring to a receptor binding residue (134, 262, Tab. 3.3). The basic residue R171 is replaced by an acidic asparagine in ghost shark's arrestin-1.1., probably impairing its function as a phosphate sensor.

Figure 3.9: Maximum likelihood tree of exon 5 sequences from arrestins of spotted gar, ghost shark and little skate. As little skate has an extremely fragmented genome, the nucleotide sequence of the longest exon, i. e. exon 5, was used to build a phylogenetic tree. In little skate, five full-length and one partial exon were detected. The maximum likelihood tree was built with `PhyML` with HKY69+I+G model (α = 1.3 and 13 % invariable sites). The four orthology groups are clearly visible. Generally, little skate sequences cluster with ghost shark sequences. Concerning the *SAG* paralogs (grey cloud), two distinct *SAG* genes exist, in ghost shark and little skate, suggesting a shared *SAG* gene duplication in the common ancestor. Exons of non-visual arrestins (green clouds) clearly cluster together, splitting in *ARRB1*s (light green) and *ARRB2*s (dark green). It is not clear, whether little skate possesses two *ARRB2*s, as exon 5 on *AESE01104697.1* is partial and the encoded part is 92.9 % identical to exon 5 on *AESE011647096.1*. The phylogenetic tree was visualized with `Dendroscope 3.5.7` (Huson and Scornavacca, 2012).

Increase of arrestin number in ray-finned fish as a consequence of 3R-WGD

As stated above, the spotted gar genome harbors four arrestin paralogs as the majority of lobe-finned fish (Fig. 3.10). Contrarily, six or seven arrestin genes are encoded in all eight investigated teleost genomes (Fig. 3.10).
Our results confirm and extend the reports of 1:many arrestin orthologs to human in medaka and zebrafish by Imanishi, Hisatomi, and Tokunaga (1999) and Renninger, Gesemann, and Neuhauss (2011). The increased number of paralogs is explained by the teleost-specific 3R-WGD that happened about 320-350 mya (Glasauer and Neuhauss, 2014). The timing roughly corresponds to the 3R-WGD happening directly after the split of the gar and teleost lineages 315 mya (Fig. A.8). The 3R-WGD potentially doubled the number of arrestin paralogs (Renninger, Gesemann, and Neuhauss, 2011). I hypothesize that one copy of *ARRB1* was lost before the divergence of Otomorpha and euteleosts during the initial 85 my after the 3R-WGD (Fig. 3.10, A.8, Sato, Hashiguchi, and Nishida (2009)). The other three pairs of copies are retained in the teleost ancestor.

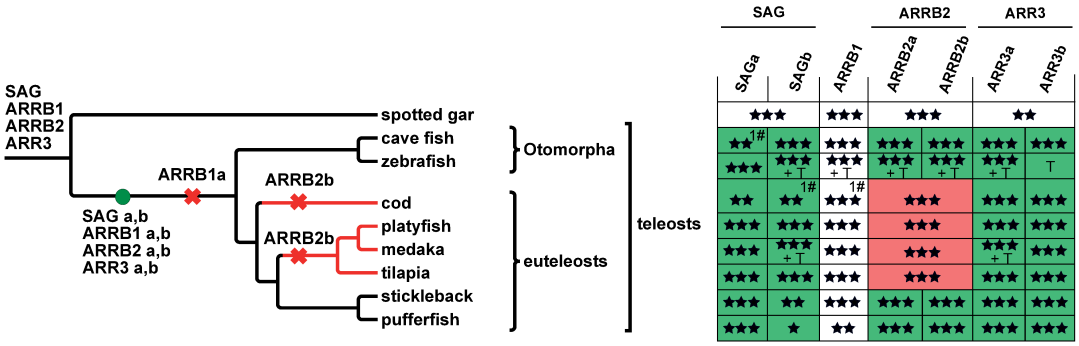

Figure 3.10: Duplication and deletion of arrestin paralogs within ray-finned fish. The teleost-specific whole genome duplication (WGD) doubles the number of arrestin paralogs in teleosts. The two resulting copies of each paralog are depicted as a and b. Zebrafish *ARR3b* was annotated in *GRCz10* as the gene was missing in the originally investigated genome version *Zv9*. The species tree was created based on (Betancur-R. et al., 2013) using `Treegraph 2.0.54` (Stover and Muller, 2010). Crosses depict gene losses. See caption of Fig. 3.7 for additional description of symbols.

While the non-visual *ARRB2a* evolved mostly under purifying selection (80 % of sites, $\omega = 0.013$), the majority of sites (98.8 %) of *ARRB2b* evolved under neutral evolution in the ancestral teleost lineage indicating a relaxation of evolutionary pressure for one of the copies. The respective residues of *ARRB2a* and *ARRB2b* are under strong purifying selection ($\omega = 0.013/0.015$) in all other branches of the teleost phylogeny where the respective paralog is encoded. Surprisingly, *ARRB2b* was lost independently along two different branches of euteleosts (Fig. 3.10) and gained comparably high divergence in the other two investigated teleosts within the order Percomorphaceae, stickleback and pufferfish (*Takifugu rubripes*). This results in a lower average protein identity of 79.3 % between Percomorphaceae *ARRB2a* and *ARRB2b* in comparison to 90 % sequence identity in the investigated two Otomorpha species. Comparison of expression patterns of *ARRB2a* and *ARRB2b* reveals that both paralogs possess not only a high sequence identity, but also a highly similar developmental and spatial expression in zebrafish (Otomorpha), while no data is available for Percomorphaceae (Fig. 3.11 A). Unsupervised MCA shows each of the two Percomorphaceae *ARRB2b*

as a separate cluster that is clearly distinct from the group of all other *ARRB2* in teleosts (*ARRB2a* and Otomorpha *ARRB2b*). Moreover, this pattern is endorsed by phylogenetic inference, where *ARRB2a*, Otomorpha *ARRB2b* and spotted gar *ARRB2* form a well supported monophyletic clade (72 % BS, 98.7 pp) with Percomorphaceae *ARRB2b* as the closest outgroup. The observed high sequence divergence between the Otomorpha and Percomorphaceae *ARRB2b* as well as within Percomorphaceae might have led to differences in functions. Interestingly, the methods `RDP` and `GENECOV` detected signals of gene conversion/recombination for a 51 nt region between stickleback *ARRB2a* and *ARRB2b*. Shared differences between the Percomorphaceae *ARRB2b* identified by manual inspection concern residues binding to IP6 (K161Q, Milano et al. (2006)), the phosphate sensor R166 (mutated to Q/H, Vishnivetskiy et al. (2011)) and AP-2 binding residues (R395G, Burtey et al. (2007)). The identity of many residues that mediate receptor specificity is conserved in stickleback *ARRB2b* in comparison to *ARRB2a*, while it differs in pufferfish *ARRB2b* (e. g. C17S, A87V, T137S, H190N, Q256H). Interestingly, homologous (R166, R395) or neighboring to homologous residues (K161, P253) are specificity determining in sea urchin *ARR0.1*.

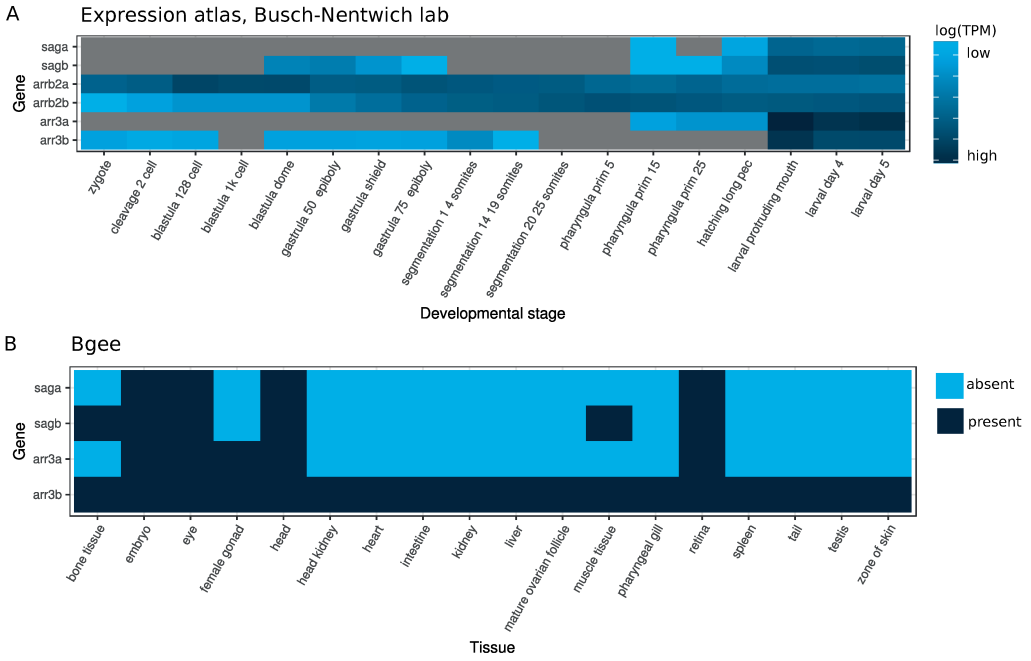

Figure 3.11: Temporal and spatial expression of arrestin genes in zebrafish. A – Temporal expression pattern during embryonic development extracted from the Expression atlas (Petryszak et al., 2016). Expression values < 0.5 TPM are shown in gray. I would like to thank the Busch-Nentwich lab for providing the RNA-seq data. B – Expression presence and absence mapped to unified anatomical structure ontology terms by `Bgee` (Bastian et al., 2008). The presence-absence pattern of *ARRB2a, b* is not shown as both paralogs are present in all structures, where information was available in sufficient quality.

The paralogs of the visual arrestins, *SAGa* and *SAGb* as well as *ARR3a* and *ARR3b* that arose during 3R-WGD (ohnologs) persisted in all investigated teleost species and evolved with similar rates since their emergence. The single orthology groups are monophyletic with $>= 80$ % BS ($>= 99.9$ % pp) and cluster together with their respective paralogy group as expected. Spotted gar *SAG* clusters near the root of the *SAGb* subgroup. *SAGa*s/*SAGb*s and *ARR3a*s/*ARR3b*s are also recognized as separate groups in unsupervised MCA applied to alignments of *SAG* and *ARR3* in

ray-finned fish, respectively, emphasizing their sequence divergence. About 17 % and 13 % of residues evolved under positive selection in the ancestral branches of *SAGa* and *SAGb*, respectively (Tab. B.6, 3.3, Fig. A.2 C). SDPs of the teleost *SAG* and *ARR3* groups overlap with phosphate sensing and receptor binding residues (Fig. 3.12 A-D). Visual inspection of the MCA revealed that *SAGb* and *ARR3b* of the teleost orders Otomorpha and euteleosts show systematic differences within their respective monophyletic groups. Otomorpha *SAGb*s form a subgroup within teleost *SAGb* that includes spotted gar *SAG* in the MCA. The subdivision in Otomorpha and euteleosts is apparent upon inspection of the low affinity IP6 binding site (Otomorpha-specific positions A164, Q165, R167, Fig. 3.12 A), but not in receptor binding residues. Within the low IP6 binding site, the positively charged residue R167, which is part of that motif, was substituted by a neutrally or negatively charged amino acid in euteleosts *SAGb* (E, Q, A, Zhuang et al. (2010)). In Otomorpha *SAGb*, all *SAGa* and *SAG* of spotted gar, the positively charged arginine is conserved. A neighboring residue (165) was converted to arginine in the teleost *SAGa* stem lineage, while this position is occupied by negatively or neutrally charged amino acids in *SAGb* (Q, C, D) with glutamine being specific for Otomorpha *SAGb*. This is further confirmed by the fact that 12 % of sites of *SAGb* evolve under positive selection in the ancestral branch leading to the sister group Acanthopterygii (euteleosts without cod, Fig. A.2 C).
As for *SAGb*s, also *ARR3b*s of Otomorpha form a subcluster within teleost *ARR3b*s in MCA. In contrast to *SAGb*, euteleost *ARR3b* differ systematically from all other teleost *ARR3* sequences with respect to receptor binding residues (e. g. positions 76, 246, 248, 254, Fig. 3.12 D). C254 was identified as one of the sites that evolved under positive selection (in total 14 %) in the ancestral branch leading to euteleosts (Tab. 3.3, B.6, Fig. A.2 D). Differences in euteleosts *ARR3b* as compared to Otomorpha *ARR3b* are also apparent in phosphate sensing residues with the latter one conserving position K157, which is occupied by a negative or hydrophobic amino acid in the first group. *ARR3a* possesses one or two additional positive charges in the same sequence stretch as compared to mammalian *ARR3* orthologs (K152 or K154 and K157, Fig. 3.12 C, Zhuang et al. (2010)). The low affinity IP6 binding site is conserved in all vertebrate *ARR3* otherwise, although IP6 binding has not yet been characterized experimentally for this paralog. Although the visual ohnologs of *SAG* and *ARR3* share expression in several anatomical structures in zebrafish (eye, retina, embryo, head), the *b*-ohnolog is expressed in tissues, where the *a*-ohnolog is absent pointing to spatial subfunctionalization (*SAGb*: muscle, bone; *ARR3b*: 13 tissues including muscle, bone, spleen, liver, see Fig. 3.11 B). Furthermore, the ohnologs possess a distinct temporal expression pattern during embryonic development in zebrafish (Fig. 3.11 A).

Loss or pseudogenization of *ARR3* in afrotherians, xenarthrans, and the common shrew

Within the second clade of bony fish, the lobe-finned fish, a single gene for each of the four paralogs is retained with a few exceptions: (1) Loss or pseudogenization of *ARR3* in afrotherians, xenarthrans and common shrew (*Sorex araneus*); (2) Retrogene formation and pseudogenization of *ARRB1* and *ARRB2* in marsupials; (3) likely loss of *ARRB2* in birds (Fig. 3.13).
The mammalian superorder Afrotheria consists of two clades, African insectivores (Afroinsectiphilia) and paenungulates. *ARR3* of African elephant, *Loxodonta africana*, and rock hyrax, *Procavia capensis* (paenungulates) are degraded to pseudogenes to different extents (Fig. 3.14). In elephant, exon–intron structure and fragments of

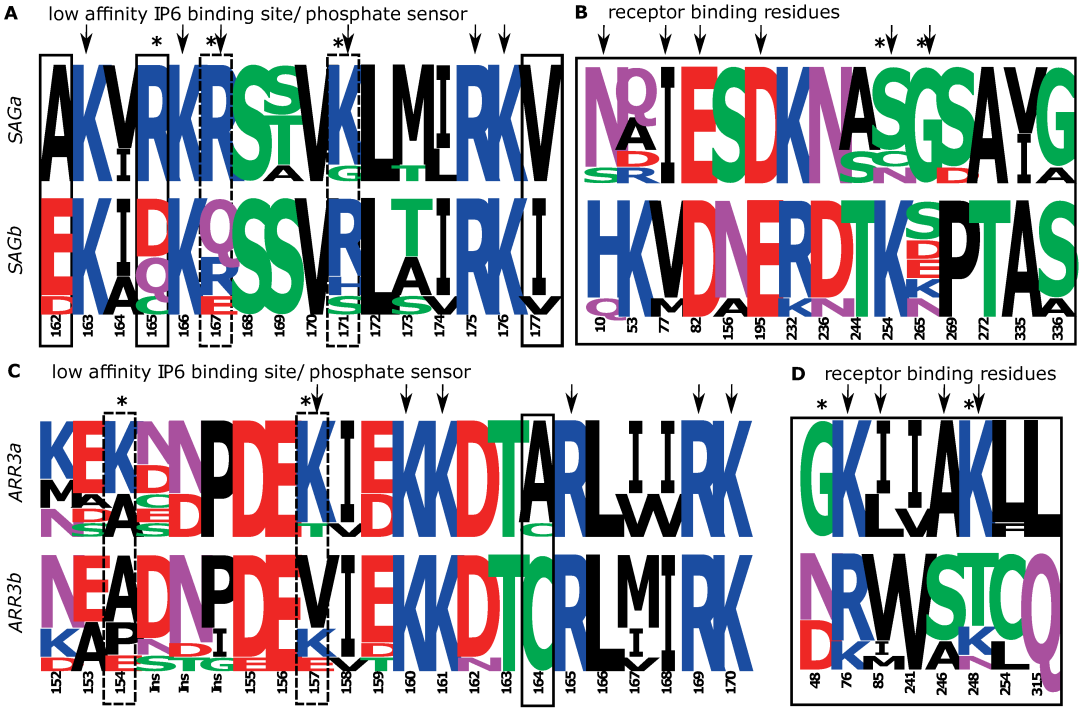

Figure 3.12: Specificity determining positions discriminating each pair of duplicated visual arrestins in teleosts. Amino acid frequency logos are shown for *SAGa* vs. *SAGb* (A, B) and for *ARR3a* vs. *ARR3b* (C, D) in teleosts. Positions that are known to directly confer a specific functionality in mammalian arrestins are marked by arrows. Of these, some mutations change the charge of the respective residue (marked with *). Positions identified by specificity determining position (SDP) analysis are highlighted by black boxes. As receptor specificity is mediated by a rather big interface, only the SDPs are shown that are known to be involved in receptor binding and their first and second order neighbors. Additional positions that show differences in both groups (manually identified) and might be associated with the respective function are highlighted with a dotted box. Please see Tab. 1.3, B.8 for literature references of the functions. The numbering refers to the position numbers in bovine *SAG* and *ARR3*, respectively. Results are also summarized in Tab. B.7. The figure was created with `Weblogo` (Crooks et al., 2004). Abbreviation: Ins – Insertion in comparison to reference.

the sequence are homologous with 61 % identity to the human arrestin ortholog as compared to at least 80 % pairwise identity for other placental *ARR3*. The upstream and downstream syntenic genes are situated on a continuous, gap-less sequence stretch. Even under assumption of non-canonical SSs, the best annotation of elephant *ARR3* encodes for six stop codons within the putative protein-coding sequence. Mutations disrupt former key functional elements, e. g. the polar core (D297Y) or residues important for receptor specificity (C282F, T259/261). In contrast to elephant, sequence in between *AWAT1* and *PDZD11* is not completely covered in the genome of hyrax (Fig. 3.14). Nevertheless, some sequence clearly shows similarity to exons 8-10, 12, 14 and 16 of *ARR3* with 26 % identity to the human *ARR3* query. Although sequence in between exons 12 and 14 is completely sequenced, exon 13 cannot be identified by homology search. Annotation attempts result in two stop codons and one frame shift. In conclusion, this points towards a degradation of *ARR3* to pseudogenes in both investigated paenungulate genomes.
In contrast, *ARR3* is completely lost in the genome of the xenarthran armadillo *Dasypus novemcinctus*. The orthologs of the syntenic genes from human, *P2RY4* and

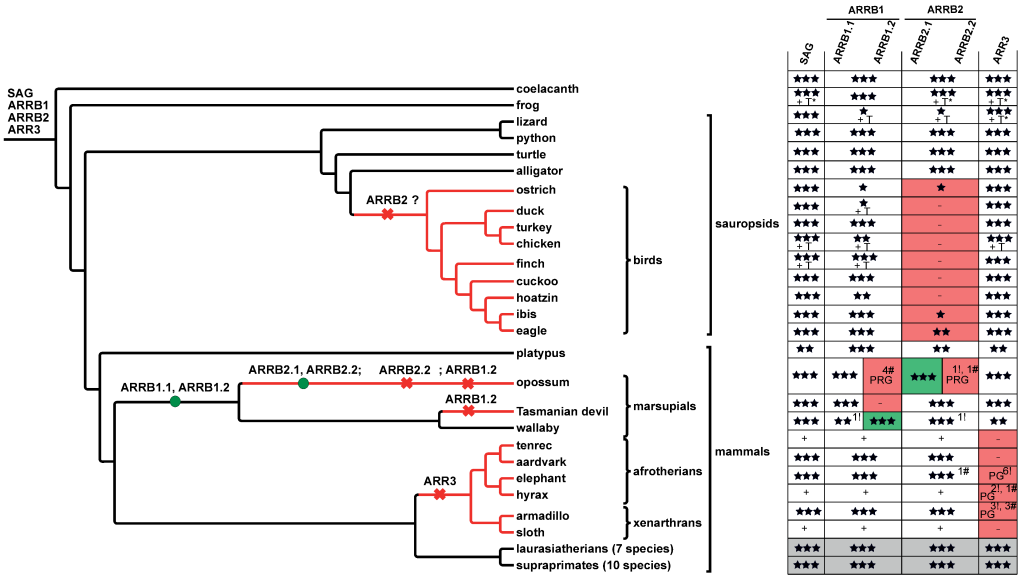

Figure 3.13: Duplication and deletion of arrestin paralogs in lobe-finned fish. *ARRB2* could not be detected in the genomes and transcriptomes of birds (see Tab. B.9 for other 41 investigated bird species). Additional omics-data was investigated for sauropsids. The gene loss/duplication pattern was simplified for the monophyletic groups highlighted in light grey (Fig. A.7). See caption of Fig. 3.7/3.10 for description of symbols. The exclamation mark (!) indicates the number of stop codons contained in the exon annotation, while plus (+) indicates that gene (parts) are encoded within the respective genome, but were not annotated in detail. Note that the order of the *ARRB1.2* and *ARRB2.2* losses is arbitrary. The phylogenetic tree was created using `Treegraph 2.0.54` (Stover and Muller, 2010). Abbreviations: PG – pseudogene; PRG – pseudo-retrogene.

PDZD11 are situated about 16.5 kb apart on the same contig as *ARR3* in armadillo, *JH563233* (Fig. 3.14). The loss of *ARR3* is supported by two facts: (1) No homologous sequence can be identified when blasting the nucleotide sequence of the elephant *ARR3* pseudogene against this locus (Fig. A.9 A); (2) The locus between *P2RY4* and *PDZD11* is extremely shortened in comparison to the length of the *ARR3* gene in other mammals (e. g. 22 kb in human). The `tblastn` search against the armadillo genome with the bovine *ARR3* as query retrieved one hit that did not overlap with other annotated arrestin loci (*E*-value 0.1), but with the novel protein-coding gene *ENSDNOT00000049106*. This gene was annotated by the `Ensembl` gene prediction pipeline and has the *arrestin_N* domain (PF00339, Fig. A.9 B). Nevertheless, the locus can be excluded as (1) The exon–intron structure is not conserved in comparison to other placental *ARR3*; (2) Annotation of a stop codon and several frame shifts would be necessary; (3) Sequence identity to *ARR3* in horse is extremely low with 36 %.
In the three other investigated xenarthran and afrotherian genomes, the neighboring genes are either situated on different genomic fragments (sloth and aardvark) or are lost to Ns (tenrec, Fig. 3.14). No hits were retrieved for *ARR3* in the genomes of aardvark, tenrec (both African insectivores) and sloth (xenarthran).
An independent degradation of *ARR3* was observed in the genome of common shrew (Fig. A.10). The respective region contains fragments similar to exons 3, 8, 10, 12 and 14 of *ARR3* of other mammals. Annotation with `ProSplign` retrieved a degraded gene that encodes for at least five stop codons within exons, has no start and stop codon. The complete sequence coverage between the fragments of exons 3-14 support an independent pseudogenization.

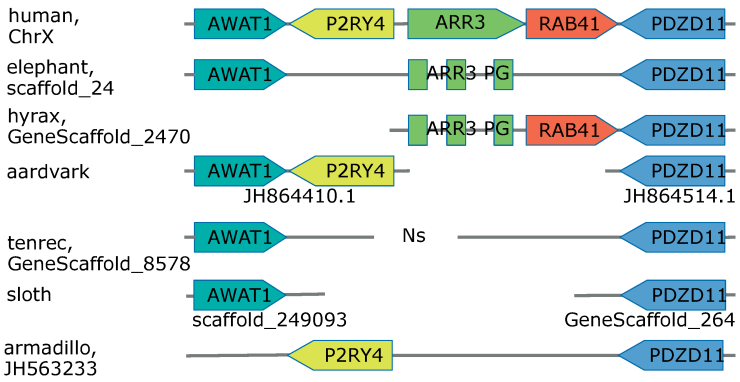

Figure 3.14: Synteny of the *ARR3* locus in afrotherians and xenarthrans. The synteny relationship of *ARR3* as known from human is conserved in all investigated mammals where the genes are situated on a continuous fragment. Within afrotherians and xenarthrans, *ARR3* is degraded into pseudogenes (PG) or lost for those species where synteny information is complete.

Retrogene formation and pseudogenization of *ARRB1*/*ARRB2* in marsupials

As another peculiarity within mammals, I identified additional *ARRB1* and *ARRB2* genes in the marsupial genomes of opossum and wallaby (*Macropus eugenii*, Fig. 3.13). Both genes, *ARRB1.2* and *ARRB2.2*, are encoded by a single exon, the main characteristic of a retrogene (section 1.2.1). The parental gene *ARRB1* possesses a much higher expression in opossum testis than *SAG* in the same species confirming the previously described expression pattern from human and mouse in a marsupial (Fig. A.11). The *ARRB1.2* retrogene seems functional in wallaby as the putative protein-coding open reading frame is intact and highly homologous to the putative parental copy and other vertebrate *ARRB1* cDNAs (Fig. A.12). The retrogene has an accelerated substitution rate in comparison to the multi-exon *ARRB1.1* as indicated by longer branch lengths (Fig. 3.5). Although a thorough analysis of SDPs with more paralogs is missing, I speculate that the phosphate sensing behavior of *ARRB1.2* is changed due to the amino acid substitutions K160E and R161Q at positions that strictly conserve the positively charged amino acids lysine and arginine in all other vertebrate *ARRB1*. In contrast, *ARRB1.2* has turned into a pseudo-retrogene in opossum indicated by four frame shift mutations within the potentially protein-coding region. Applying the parsimonious principle, I assume that a processed *ARRB1*-mRNA was inserted into the nuclear genome of the common ancestor of both species between 82-177 mya before split of Didelphimorphia and Australidelphia (Nilsson et al. (2010), Fig. A.8). Remarkably, the *ARRB1.2* retrogenes from both species share high conservation with the putative 5′ untranslated region (UTR) as annotated by `Ensembl` for the wallaby multi-exon *ARRB1.1* (Fig. A.12). The Tasmanian devil (*Sarcophilus harisii*), the third investigated marsupialian species, has completely lost the *ARRB1* retrogene.
Independently, an *ARRB2* retrogene was inserted within a cluster of zinc-finger transcription factors on chromosome 3 in the lineage leading to opossum. However, the retrogene turned into a pseudogene containing a premature stop codon and an insertion resulting in a frame-shift mutation (Fig. 3.13).

Possible loss of *ARRB2* in birds

Surprisingly, hardly any fragments of *ARRB2* were detected in bird genomes or lizard, while the respective ortholog was easily detectable in the genomes of other

sauropsids, e. g. alligator (*Alligator mississippiensis*), turtle (*Pelodiscus sinensis*) and python (*Python molurus*). This raised the possibility of a loss of the *ARRB2* gene within these species. Extensive homology search in 50 bird genomes retrieved only five species that harbor two or more complete exons of this 15 exon gene *ARRB2* (bald eagle, ibis (*Nipponia nippon*), ostrich, kiwi, golden eagle, Tab. B.9). All detected exons have a high sequence identity to orthologous exons in turtle (on average 91.3 %, at least 83.9 %). The potential loss was further tested by (1) investigating genomic synteny of *ARRB2* and (2) expression of *ARRB2* in transcriptome/EST data. First, syntenic information had to be inferred from the *ARRB2* locus in other species. Synteny information from mammals, sauropsids and coelacanth supported the conservation of the gene neighborhood with *Med11* oriented head to tail and *Pelp1* head to head of *ARRB2*, respectively, when synteny information was available. Collectively, this information suggests that *ARRB2* was located between *Med11* and *Pelp1* in the last common ancestor of lobe-finned fish and that this linkage is conserved throughout lobe-finned fish. In this study, only *ARRB2* in frog was found to have the gene *DDX27* as a neighbor in place of *Med11*. The latter was found in a completely different gene neighborhood, which might be the result of an amphibian-specific rearrangement. Nevertheless, none of the potential neighboring genes, *Med11* or *Pelp1*, was detected in the genomes of the investigated bird species or in lizard.
Second, the genome-focused approach was complemented using specific bird transcriptome data sets (section 3.2.3). Whole or partial hits were retrieved for *SAG*, *ARRB1* or *ARR3*, while in general no hits were retrieved for *ARRB2*. Within the investigated chicken ovary expression data (Boardman et al., 2002), some fragments were recovered that could not be assigned to neither *ARRB1* nor *ARRB2* unambiguously, but were similar to a non-visual arrestin. Neither of the two strategies provided evidence to reject the hypothesis that *ARRB2* has been lost in birds. In contrast, a query of the `NCBI EST` database retrieved both non-visual arrestin transcripts in lizard confirming the integrity of the *ARRB2* gene in reptiles.

3.3.4 Evolution of arrestin functional elements

Loss and gain of functional elements

Scanning the `Pfam 28.0` database using `hmmscan` confirmed that more than 95 % of all annotated deuterostome arrestins possess an *arrestin_C* and an *arrestin_N* domain. For the few other arrestins, sequence data was missing in the respective region. Apart from the *arrestin_C* and *arrestin_N* domains, the following other domains were detected in more than 25 % of the deuterostome arrestins: *BatD*, a membrane spanning protein connected to oxygen tolerance in bacteria, the clathrin-adapter complex 3 beta 1 subunit C terminal domain (*AP3B1_C*) and the arrestin-N terminal like domain (*LDB19*), which belongs to the arrestin N-like clan (Fig. A.13). The domains were not specific for certain orthology groups. For *AP3B1_C*, all obtained hits had an *E*-value < 0.014 (conditional *E*-value $<$ 9.4e-05) and covered 19-47 % of the profile. Mapped onto arrestins, the domain overlapped with the beginning of the *arrestin_C* domain and covered residues that are known to be involved in receptor, IP6 and phosphodiesterase binding (residues 192-237 in bovine *ARRB1*). AP3B1 is part of the adapter protein-complex and interacts with clathrin as well as with accessory proteins.
As expected, known key functional motifs such as the phosphate sensing residues (Gurevich et al., 2014), the polar core residues (Hirsch et al., 1999), the residues involved in the three element interaction, the sequence of the receptor-binding finger

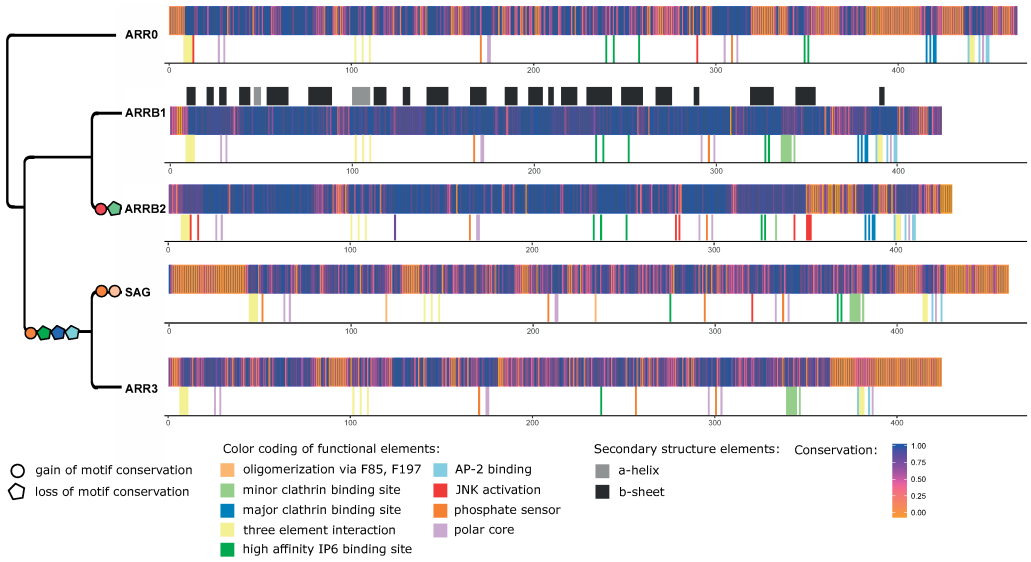

Figure 3.15: Changes in conservation patterns and functional motifs of arrestins. Conservation of alignment positions of the individual orthology groups is shown. The conservation was calculated according to the Method of Karlin (Karlin and Brocchieri, 1996) using `AACon` (Manning, Jefferson, and Barton, 2008) for each orthology group separately. Sequences with a coverage < 90 % as well as all lamprey sequences were excluded. Functional motifs characterized in one or several paralogs were projected onto the individual alignments solely based on sequence homology. Putative loss (pentagon) and gain (circle) events based on conservation of the respective motifs were projected onto a simplified arrestin gene tree. The order of motif gain and loss on the respective branch was chosen arbitrarily. Positions were not marked if they did not conserve the amino acid known to be part of the motif in that orthology group in a majority of representatives. Some positions are marked although their conservation is restricted to a phylogenetic group as indicated by their lower conservation score (e. g. oligomerization is specific for lobe-finned fish *SAG*). The secondary structure based on the crystal structure of 1G4R (Fig. 1.9) is mapped onto the alignment of *ARRB1*. Note that only a selection of known motifs are shown.

loop (Szczepek et al., 2014) and the low affinity IP6 binding site are conserved in all *ARR0* and vertebrate arrestins (Fig. 3.15). Exceptions are pointed out in the respective sections. The great majority of residues of all arrestins evolved under strong purifying selection and are highly conserved. However, recently duplicated paralogs can behave differently in respect to conservation and selection. Surprisingly, *ARR0*s do not conserve the minor clathrin binding site (CBS) indicating that this motif was probably acquired shortly before the emergence of the four vertebrate arrestins or that the sequence diverged so that it cannot be detected with the applied E-value settings. The loss is further supported by the absence of the respective motif in *ARR0* protein and transcript evidence available for three different species (Fig. 3.7). I propose that the majority of functional innovations arose due to the duplication of *ARR0* as they are commonly conserved in the respective orthology group in vertebrates. For example, arrestin-3 binds and activates JNK3, while arrestin-2 does not (McDonald et al., 2000; Song et al., 2009; Seo et al., 2011). The residues S13 and C17 previously identified to mediate JNK3 binding and activation (Zhan et al., 2016) are strictly conserved in all *ARRB2* except for lamprey and *ARRB2b* pufferfish, while being different in *ARRB1* (Fig. 3.15). *ARR0* shares residues with both non-visual arrestins – those that are conserved across *ARR0*, *ARRB1* and *ARRB2*, but also conserves paralog-specific

residues within the N-terminal 25 residues (Seo et al., 2011). The conservation of most other positions known to mediate JNK activation is restricted to a phylogenetic group of *ARRB2* such as conservation of the sequence stretch H350D351H352 in mammals and of L278xS280 in lobe-finned fish, respectively. An exception is position V343 in the *arrestin_C* domain of arrestin, which is conserved in all *ARRB2* except Otomorpha *ARRB2a*. Interestingly, all sea urchin *ARR0.1* sequences carry a conserved valine here, while all other *ARR0* carry threonine at the homologous position, which is characteristic for arrestin-2.
In both visual arrestins, the high affinity IP6 binding site, the AP-2 binding site, the major CBS and the first PxxP motif involved in binding of the kinase c-Src are not or are loosely conserved, in contrast to non-visual arrestins (Fig. 3.15). *SAG* and *ARR3* generally conserve the key residues K163, K166 and K167 (K157, K160, K161, respectively) of the low affinity IP6 binding site with the exception of the teleost *b*-ohnologs arising from the 3R-WGD (Fig. 3.12). Other key mutations that occurred in visual arrestins in comparison to *ARR0* involve A253D, which was hypothesized to weaken the hydrogen bond network of the pre-activated state in comparison to non-visual arrestins (Kim et al., 2013). An additional phosphate binding residue, R18 (Sutton et al., 2005), is conserved in all *SAG* sequences. The residues F85 and F197, which are known to be involved in oligomerization of *SAG* (Hanson et al., 2008) are strictly conserved in *SAG* of the lobe-finned fish. The C-terminus of teleost *ARR3* is shorter than in *ARR3* of other vertebrates. For example, the C-terminus of *ARR3a* and *ARR3b* in zebrafish is 31 and 24 aa, respectively, shorter than the C-terminus of *ARR3* in spotted gar. The residues missing in zebrafish are known to be responsible for the three element interaction, AP-2 binding and contribute an arginine to the polar core (Aubry and Klein (2013), Fig. 3.15, 3.16). Interestingly, the very last 10-20 aa of the C-terminus following the AP-2 binding site and the three element interaction, differ systematically among sauropsids, rodents and non-rodent mammals.
PTMs of non-visual arrestins have effects on their interactions with partners, e. g. receptors, kinases and components of the endocytosis machinery as shown experimentally (section 1.1.4). Those positions are frequently conserved within but not across orthology groups (Tab. B.10, B.11). Phosphorylation of S412 of *ARRB1* regulates clathrin binding and endocytosis (Lin et al., 1997); phosphorylation of S/T360 in *ARRB2* regulates clathrin-mediated internalization (Lin et al., 2002); nitrolysation of C409 in *ARRB2* promotes binding to clathrin and AP-2 (Ozawa et al., 2008). Other positions known to be phosphorylated and involved in the interaction with μ2 adaptin (Y54 in *ARRB1*) or the regulation of receptor binding (T178), respectively, are clade-specific and, thus, represent recent evolutionary innovations. In contrast, known ubiquitination and SUMOylation sites are conserved across orthology groups. The majority of sites emerged in the common ancestor of non-visual arrestins consistent with their need to regulate receptor trafficking and internalization more specifically (Shenoy et al., 2001; Girnita et al., 2007; Wyatt et al., 2011; Jean-Charles, Rajiv, and Shenoy, 2016). Additionally, I uncovered potential functions and conservation patterns of post-translational modified residues by overlapping the functional annotation with different phosphoproteome data sets (Tab. B.10, B.11, section 3.2.7). Five of those positions (T254, Y258, Y47/Y48, S194, S267/S268), among them all three PTMs of *SAG*, are characterized as receptor SDPs or are situated next to them. Modifications at positions T374, T404 and T410 might influence the binding of clathrin or AP-2 as do other proximal residues. Finally, phosphorylation of T173 and K178 may regulate the binding of c-Src via the first and functionally characterized PxxP motif.

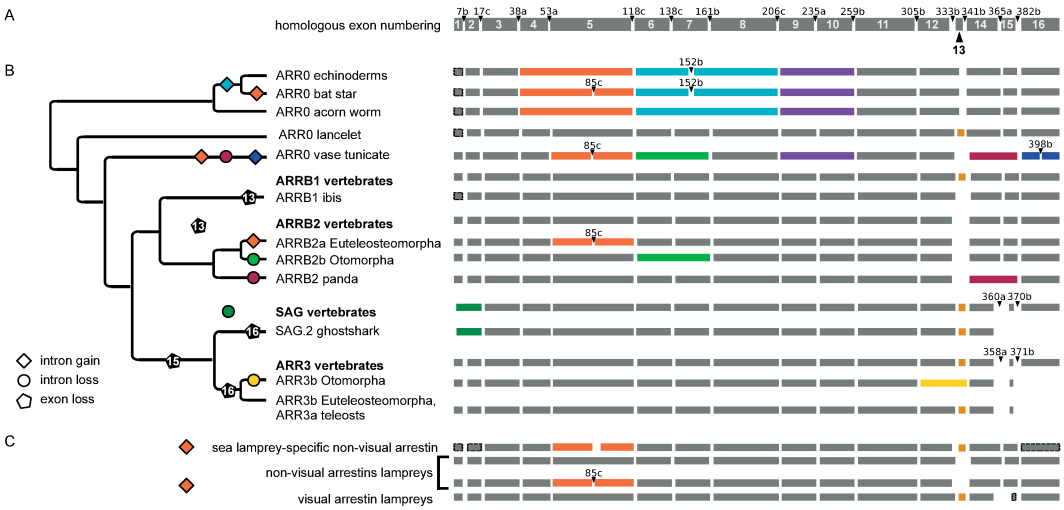

Figure 3.16: Evolutionary changes in exon–intron structure of arrestins. A – exon–intron structure of the bovine *ARRB1* gene. Exon and intron numbering is imposed onto arrestin homologs by sequence alignment. Positions of introns refer to their position on the amino acid sequence of cow arrestin-2 with a-c indicating their position after the first, second or third base of the codon, respectively. B – exon–intron structure of arrestins (right hand side) is associated with a simplified gene tree (left hand side). Exons are shown as grey and colorful boxes, whereby homologous regions are "aligned" below each other. Colored exons highlight differences in exon–intron structure (intron gain, intron loss, exon loss). Changes in intron positions in comparison to the reference amino acid sequence of cow arrestin-2 are given whenever deviating except for the positions surrounding exons 13 and 15, which occasionally deviated by few nucleotides in our annotation. Information about the corresponding exons was not available in the genomes if boxes are surrounded by a dotted line, but are assumed to be the same as in the 1:1 ortholog of the closest relative. If an unequivocal scenario of intron loss or gain is in accordance with the maximum parsimony principle, these events are indicated in the phylogenetic tree. Paralogs of species that share the exon–intron structure are summarized to phylogenetic clades, e. g. *ARRB1* vertebrates. Structural differences in comparison to the family are shown right below associated with the corresponding species or phylogenetic clade. Losses of coding sequence (exons) are indicated by black pentagons with respective exons given as a number in the pentagon. The phylogenetic tree was created using `Treegraph 2.0.54` (Stover and Muller, 2010). C – exon–intron structure of lamprey arrestins. Note that the length of the exon boxes is drawn to scale.

Hotspot of exon gain/loss at positions determining receptor specificity

The exon–intron structure of the vertebrate arrestin paralogs is highly conserved, preserving the majority of exon–intron boundaries of their last common ancestor, *ARR0* (Fig. 3.16). Nevertheless, changes in gene structure including loss of coding sequence, intron gain or loss are much more frequent in the arrestin gene family than in other vertebrate gene families (Ragg et al., 2009). The intron gain/loss events were reconstructed according to the maximum parsimony principle. Most changes in exon–intron structure are caused by single events with the notable exception of hotspots at positions 85c (five independent intron gains), at 138c (three independent events), 235a (two independent events) and 365a (two independent intron losses, Fig. 3.16 B). In accordance with the propensity for these events in paralogous gene families as discussed by Babenko et al. (2004) and Roy and Penny (2007), these gene

structure changes mainly occurred within arrestin genes that underwent a tandem duplication (exemplified by loss of exon 16 in *SAG* of ghost shark) or WGDs (loss of exon 16 in *ARR3* of teleosts, gain of intron 85c in *ARRB2a* of euteleosts, loss of intron 138c in *ARRB2b* and of intron 333b in *ARR3b* of Otomorpha). This can be further illustrated by to the emergence of the four arrestin paralogs by 2R-WGD from *ARR0* accompanied by at least one intron loss event (intron 7b) in *SAG* and a loss of coding sequence in the ancestor of *SAG* and *ARR3*, as well as in *ARRB2* (exons 15 and 13, respectively, Fig. 3.16 B). Interestingly, I observed the gain of intron 85c between 148-230 mya in the ancestor of euteleosts, a branch of teleosts, for which frequent intron gains were described previously for several GPCRs and the serpin gene family (Ragg et al., 2009; Kumar et al., 2011; Kumar, 2015).
Surprisingly, introns were gained five times independently at position 85c of deuterostome arrestins (Fig. 3.16 B/C). Four of these events occurred at the exact same position, while the exact position of intron gain in the sea lamprey-specific non-visual arrestin cannot be resolved with the available data. This paralog is excluded from the following conclusions. Two of those intron gains occurred within vertebrates (lampreys and euteleosts), a very rare event for this clade (Ragg et al., 2009; Coulombe-Huntington and Majewski, 2007). Introns are known to preferentially insert into DNA sequences that carry an upstream AG and a downstream G in respect to the insertion site. This site, "AG|intron|GY", has been termed proto-splice site in literature (Sverdlov et al., 2004), whereby | denotes a SS. The identity of thymine as 'Y' is strongly preferred (section 1.1.4). Alignment of the intron-containing paralogs with their intron-deficient orthologs of closely related species revealed a prevalence of intron gain at this position caused by the existence of a proto-splice site in all intron-containing paralogs (Fig. 3.17). Newly gained intron sequences of the respective arrestin paralogs differed in length (minimally 87 nt in tilapia (*Oreochromis niloticus*), maximally 1,358 nt in vase tunicate) and did not have any apparent sequence homology to the nuclear or mitochondrial genome of the same species or the respective intron sequences of other species omitting the inference of the origin of this intron.

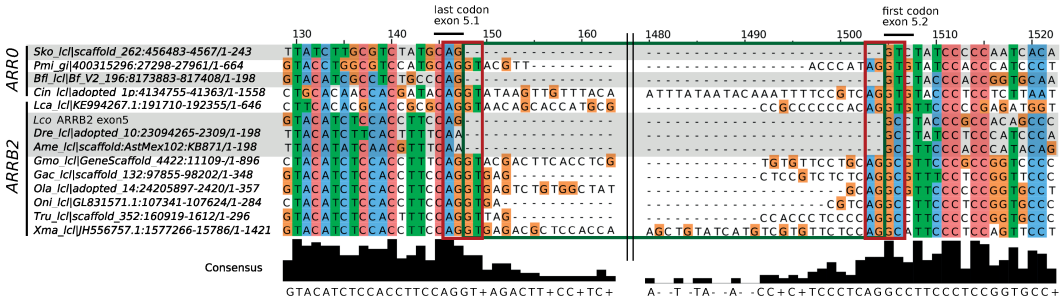

Figure 3.17: Alignment of exon–intron borders after insertion of intron 85c into exon 5. Intron 85c is found in *ARR0* of bat star (Pmi) and vase tunicate (Cin), but not in acorn worm (Sko) or lancelet (Bfl, highlighted in grey). Exon 5 of one of the non-visual arrestins in lampreys (shown: arctic lamprey, Lca) as well as in *ARRB2* in all euteleosts (Gmo, Gac, Ola, Oni, Tru, Xma) is split into two parts, denoted as 5.1 and 5.2. In contrast, exon 5 of *ARRB2* is not split in Otomorpha (Dre, Ame) and spotted gar (Lco, grey). Only the 5'- and 3'-parts of the intron sequences are shown (green box), while the larger inner region is left out being non-informative (black lines). The proto-splice site motif 'AGGY' (red boxes) is conserved for all species genes shown except for Otomorpha ('AAGC'). The alignment was visualized with `Jalview 2.8.1` (Waterhouse et al., 2009).

No codon spans exons 5.1 and 5.2, the first and the second part of exon 5, respectively. The last codon of exon 5.1, CAG, is translated into glutamine, which is conserved in

all but two inspected arrestins (Fig. 3.17). The first codon of exon 5.2 is much less conserved translating into different non-polar, aliphatic amino acids (with descending frequency: V, I, L, M) in visual arrestins (V90 in *SAG*, V85 in *ARR3*). The thymine at the second codon position is thus conserved in all species except for three, which encode the amino acid alanine. In non-visual arrestins, the same codon translates into small amino acids (A, S) due to the conservation of cytosine at the second codon position (S86 in *ARRB1*, A87 in *ARRB2*) except for five paralogs encoding thymine. Interestingly, one of those exceptions is the first codon of exon 5.2 in the putative *ARRB2* of lampreys (GT[ACTG]), which encodes valine (Fig. 3.17). This codon identity might have been the prerequisite for insertion of an intron at position 85c in the lamprey ancestor as it is part of the proto-splice site pattern.
Apart from intron gains at this position in non-visual arrestins in the ancestor of *ARRB2a* in euteleosts and in lamprey *ARRB2*, introns within exon 5 are also observed in *ARR0* of vase tunicate and bat star. All *ARR0* conserve the proto-splice site "AG|GT" tolerating an amino acid with a voluminous side-chain (valine) at this position (Fig. 3.17). Interestingly, V90 in bovine arrestin-1 is not surface-exposed. It is located between the two β-sheets of the *arrestin_N* domain, making contacts with several other hydrophobic residues (Han et al., 2001). Its substitution with a small side chain residue characteristic for non-visual arrestins (A or S) enables arrestin-1 binding to non-cognate M2 muscarinic receptor (Han et al., 2001). Therefore, a large hydrophobic residue in this position likely makes the *arrestin_N* domain more rigid, predisposing an arrestin to be more GPCR subtype-specific (Vishnivetskiy et al., 2011; Gimenez et al., 2012).

Evolution of the clathrin–arrestin interaction

The two CBS represent functional key motifs of arrestins and are encoded by single exons (exons 13 and 15 in the longest isoform of *ARRB1*). Omission of exon 13 during splicing results in a protein that mediates receptor endocytosis less efficiently than full-length arrestin-2 (Kang et al., 2009) thus representing a mechanism to regulate binding to the endocytosis machinery (section 1.5.3). Sequence coding for the minor CBS is completely missing in the highly similar *ARRB2* paralog (Fig. 3.16).
In *ARRB1* as well as in the visual arrestins, the number of nucleotides coding for this exon is strictly conserved, thus maintaining the reading frame if the corresponding exon was spliced out (Tab. B.12). Interestingly, the minor CBS, encoded by exon 13, is conserved in *ARR3* following the consensus motif in all clades or contains conservative mutations (Fig. 3.18). The sequence observed in mammalian and bird *SAG* deviates in one and two positions from the consensus motif, respectively, suggesting a decreased affinity to clathrin. In fish *SAG*, the minor CBS is severely shortened probably resulting in a loss of function (Fig. 3.18). Within Otomorpha, the intron between exon 13 and exon 14 is lost omitting selective exclusion of the minor CBS by alternative splicing. The homologous exon is missing in all *ARR0*, which possess the major, but not the minor CBS. Under assumption of a visual and a non-visual proto-arrestin, the most parsimonious scenario is the gain of the minor CBS before 1R-WGD resulting in an advanced *ARR0* with two CBSs similar to *ARRB1*. After 1R-WGD, the major CBS (exon 15) was lost in the visual proto-arrestin due to an extreme shortening of exon 15 to 10–16 nt (Fig. 3.16 B). In contrast, both CBSs persisted in the non-visual proto-arrestin with the minor CBS (exon 13) been lost after 2R-WGD in *ARRB2* (Fig. 3.16 B).
The AP-2 binding site is encoded by exon 16 and completely conserved across non-visual arrestins (Fig. 3.19). Residues 385–387 within the motif are part of the three

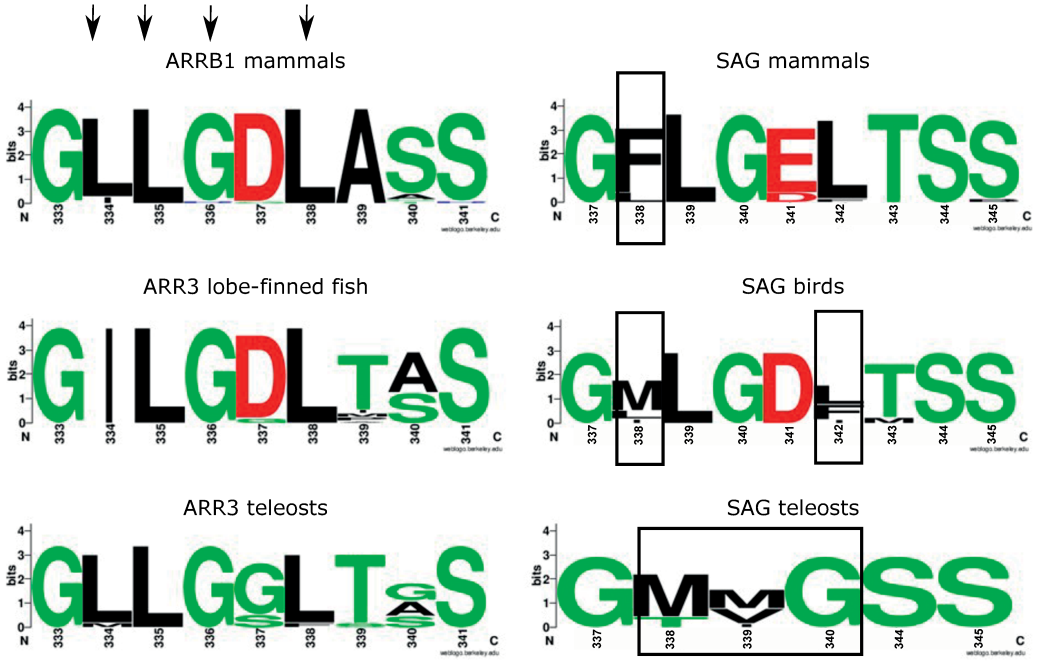

Figure 3.18: Conservation of the minor clathrin binding site in arrestins across deuterostome evolution. Sequence logos are shown for specific deuterostome clades. The respective motifs have been characterized in mammalian *ARRB1* (functional sites are marked by arrows). Deviations from the consensus motif are marked by boxes.

element interaction, one of the key motifs for arrestin activation. As in the case of the CBS, both visual arrestins experienced mutations in the consensus motif that likely decreased their affinity to AP-2. This effect might be more pronounced in placental mammals, where two of the five consensus amino acids are mutated to an amino acid with different biochemical properties (Fig. 3.19).

Conservation of possible isoforms

Apart from sequence conservation and gain of coding sequence, the expression of splice variants determines protein function. As numerous splice variants are reported for arrestins considering different sources, I restricted the assessment of the genetic prerequisites for the conservation of splice variants to those isoforms that are consistently reported for different paralogs (section 3.2.7). In general, diversity of isoforms is higher for non-visual arrestins than for visual arrestins. The annotation of splice variants confirms that the isoform used as query for homology search for each paralog is expressed in the three considered species, human, cow and mouse. In contrast, the vast majority of arrestin splice variants seems to be species- and paralog-specific, given current data. Only two splicing events – skipping of exons 4 and 13 – are supported by `Ensembl` annotations in more than one species for the same paralog. Expression of homologous isoforms across orthology groups additionally points to the conservation and importance of a late translation start within exon 8, skipping of exon 12 and the p44 splice variant with an extremely shortened exon 16. In fact, all respective reading frames of cassette exons and start codons enumerated above are conserved across deuterostomes with few exceptions (Tab. B.2). Interestingly, exon 4 is fused to exon 5 in all echinoderms and hemichordates precluding skipping

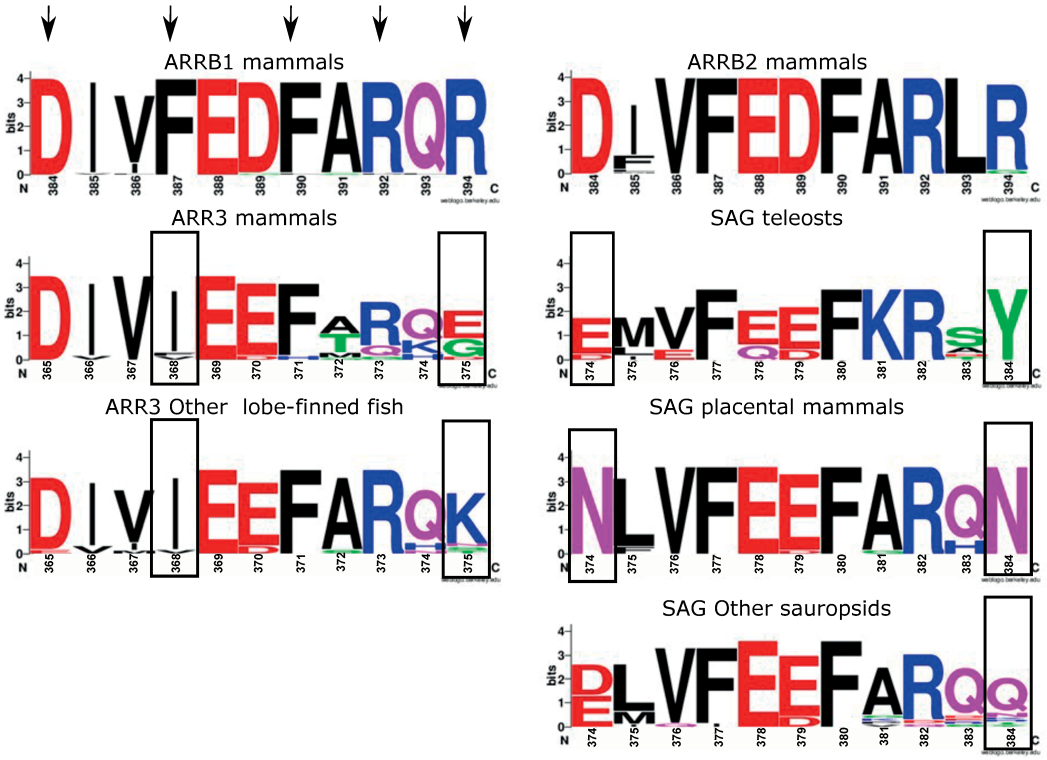

Figure 3.19: Conservation of the AP-2 motif in arrestins across deuterostome evolution. Sequence logos are shown for specific deuterostome clades. The respective motifs have been characterized in mammalian *ARRB1* (functional sites are marked by arrows). Deviations from the consensus motif are marked by boxes.

of this exon in this clade. For example, I also investigated the conservation of two other splice variants that were only found to be expressed in human *ARR3* and *SAG*, respectively: skipping of exon 15 and a premature translation stop arising from the elongation of exon 7 into intron 161b towards an encoded stop codon. The respective reading frame and premature stop codon is conserved in the respective paralog and in specific clades of other paralogs, but not across all orthology groups (Tab. B.12). Three of the four isoforms that are conserved across paralogs are not functionally characterized. The forth isoform, skipping of exon 13, which corresponds to the minor CBS, is discussed above and has been experimentally studied previously. Therefore, I can just speculate about the possible implications of the other three isoforms based on the crystal structures of the respective paralogs. Translation start within exon 8 produces a protein that is entirely compose of the *arrestin_C* domain with one β-strand missing in comparison to the full-length domain (orange in Fig. 3.20). Nevertheless, the putative structure with seven anti-parallel β-strands organized in two sheets is consistent with the definition of the immunoglobulin-like β-sandwich (Andreeva et al., 2014).

Skipping of exon 12 ablates the high affinity IP6 binding site (Fig. 3.21) and probably affects receptor binding. Residues of exon 12 participate in both functions with several residues situated within 5 Å of the respective interaction partner in the crystal structures (Fig. 3.21 B). Exon 4 encodes a β-strand and loop region in the *arrestin_N* domain. It also possesses several residues contributing to receptor binding and specificity, although those residues are not situated in such particularly close proximity to the receptor interface in the crystal structure (within 15 Å, Fig. 3.21 A). Skipping of exon 4 and 12 results in the loss of one β-strand and a loop region in

the *arrestin_N* and *arrestin_C* domain, respectively. It is expected that all three splice variants have a major impact on the overall arrestin fold.

3.4 Discussion

This section discusses the possible biological implications and interpretations of the paralog absence, presence and divergence patterns of arrestins in deuterostomes. Arrestins are one of many gene families that duplicated during the 2R- and 3R-WGDs, which necessitates a consideration of interaction partners to set the ohnolog retention pattern into context. Results and discussion of this Chapter demonstrate how the detailed evolutionary investigation of the evolution of a single gene family based on genomic data can contribute to its functional understanding that is usually gained employing experimental methods.

3.4.1 Limitation of arrestin database annotations

In this Chapter, I employed two complementary strategies to study the evolution of arrestins. In a classical and commonly conducted database scan, I extracted annotations from `UniProtKB` and `OrthoDB` to get an overview of the evolutionary relationship of arrestins and closely related proteins, referred to as the arrestin fold family. I defined the arrestin fold family as covering the orthology groups of arrestins, *ARRDC1*, *ARRDC2/ARRDC3/ARRDC4/TXNIP* and *ARRDC5* among a small number of other proteins. I confirmed previous results by Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014), stating that the arrestin fold family predates opisthokonts with very few representatives outside of eukaryotes. Arrestins were identified in choanoflagellates and Filasterea and thus predate the emergence of animals, but do not occur in fungi. Additionally and as a main focus of this Chapter, I investigated the evolution of arrestins in deuterostomes in detail and updated existing annotations considering fragmentation of arrestin genes by employing the EMS-pipeline. As expected, arrestin

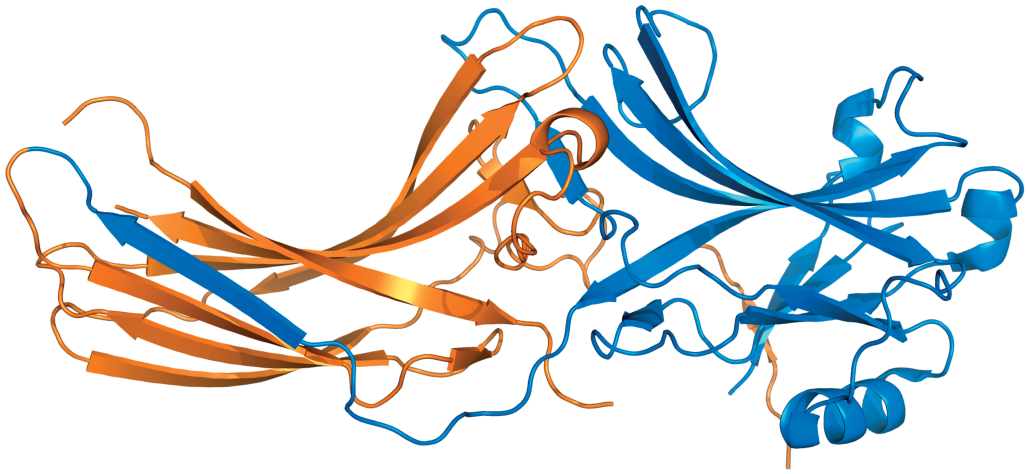

Figure 3.20: Putative effect of arrestin truncation by use of an alternative start codon within exon 8. Crystal structure of bovine *ARRB1* (PDB: 1G4R). Translation start at the conserved methionine within exon 8 results in truncation of arrestin with the orange part being present in the respective isoform. The isoform almost completely covers the *arrestin_C* domain, while the *arrestin_N* domain is not present.

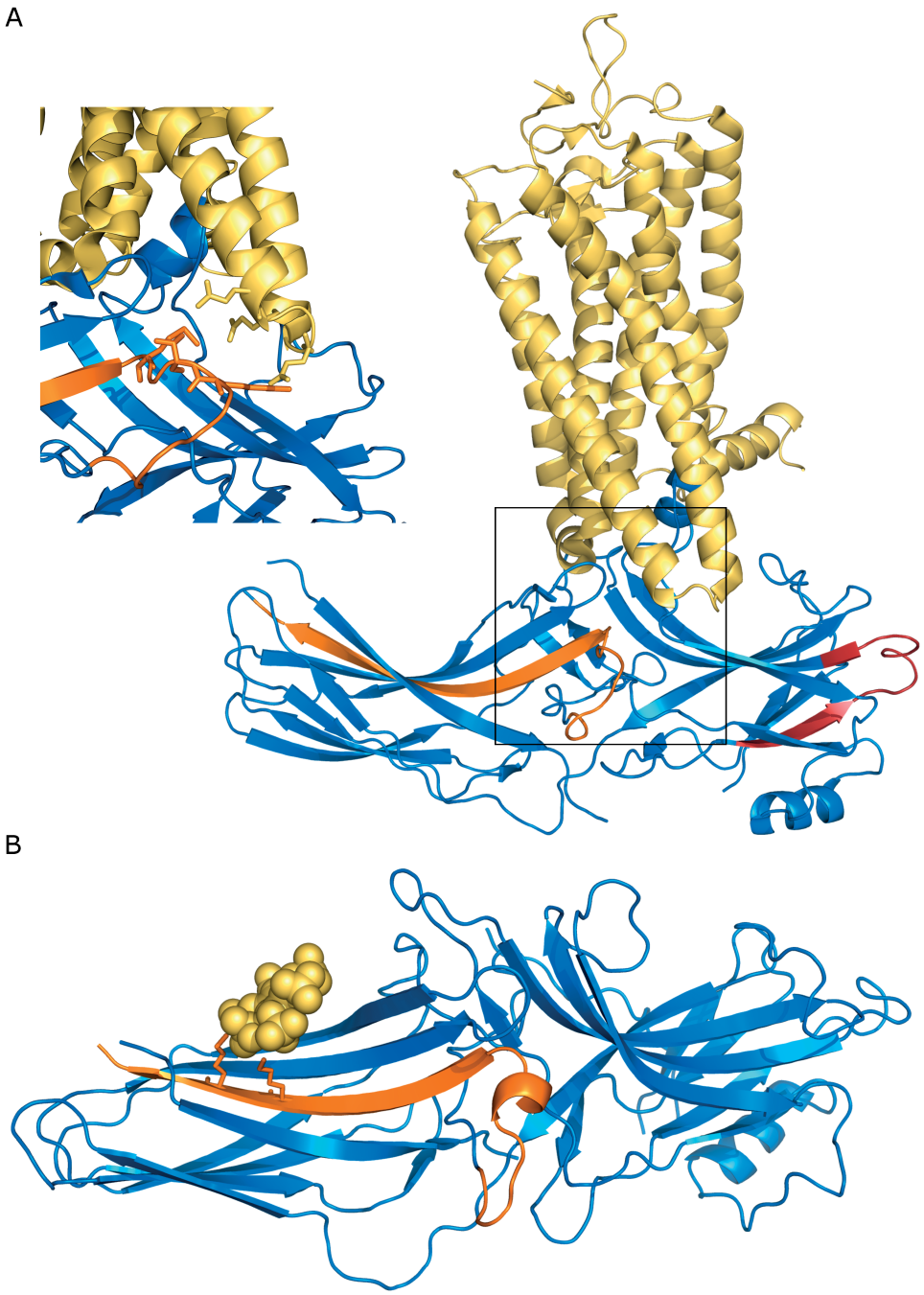

Figure 3.21: Putative effects of conserved splice variants on structure and binding interfaces of arrestins. A – Crystal structure of bovine *SAG* (blue) bound to rhodopsin (yellow, PDB: 5DGY). Arrestin splice variants with skipping of exons 12 (orange) or 4 (red) are conserved across deuterostomes. Residues within the loop encoded by exon 12 are in proximity of 5 Å to receptor residues (inlet with residues shown as sticks). B – Crystal structure of bovine *ARRB1* (blue) with inositol-hexaphosphate (IP6, yellow, PDB: 1ZSH). Exon 12 is colored in orange. Two residues are in 5 Å proximity of the ligand IP6.

annotations extracted for the arrestin orthology group as defined by `OrthoDB` and based on the pHMM scan against `UniProtKB` have severe limitations in comparison to the updated arrestin annotations. This comparison clearly demonstrates that the database scan and the resources queried therein are incomplete. Further limitations

of the database scan concern the isoform filtering of the UniProtKB data set. Filtering based on an identity cut-off in an attempt to exclude isoforms encoded by the same gene removes recent and very similar paralogs. This commonly applied strategy introducing an additional error for the estimate of paralog copies. Although the individual protein sequences extracted from UniProtKB usually have a higher quality than annotations from whole genome projects and are often experimentally confirmed, especially those databases are incomplete with a bias towards well investigated species and protein families. In comparison to UniProtKB, OrthoDB has several advantages, i. e. it contains orthology assignments and is based on genome annotations. Nevertheless, even those orthology assignments are not clean, as exemplified by the amniote-specific *ARRDC5* orthology group, which contains two invertebrate arrestin fold members (section 3.3.1). The inclusion of the distantly related members of the arrestin fold family further influences phylogenetic inference as those distant homologs cause artifacts in tree topology due to long branch attraction. For this reason, paralog counts as well as the tree of the arrestin fold family represent an approximation and details have to be interpreted with caution. These limitations exemplify the need for highly curated sets of paralogous genes in general and for arrestins in particular. The updated arrestin annotation represents one of these very rare instances and is thus ideal for evaluation of gene annotation and orthology prediction tools. The more accurate resolution of the arrestin orthology relationship in conjunction with literature mining enables me to draw conclusions about possible functional changes after arrestin duplications and deletions.

3.4.2 Arrestins as key interaction partners in a remodeling signaling network in early vertebrate evolution

The arrestin gene history cannot resolve the exact placement of the vertebrate 2R-WGD

I demonstrated here that in fact two consecutive duplications in early vertebrate evolution (presumably the 2R-WGD) led to the emergence of the four arrestin paralogs from a prototypical arrestin apparently similar to *ARR0* in vase tunicate in accordance to Nakagawa et al. (2002) and Larhammar, Nordström, and Larsson (2009). The arrestin gene tree topology reported in this Chapter supports the existence of a visual and a non-visual proto-arrestin. Nevertheless, the exact timing of the duplications could not be resolved even when mining several lamprey genomes. This is not surprising as the exact timing of the 2R-WGD is highly debated in the community without any consensus even under consideration of multiple gene families (Soltis and Soltis, 2012). Some studies place the 2R-WGD after the split of jawless fish and jawed vertebrates suggesting independent duplications in the lamprey lineage (Fried, Prohaska, and Stadler, 2003; Mehta et al., 2013), other studies argue that both 2R-WGDs took place at the root of the vertebrate tree, followed by an immediate split of both groups (Kuraku, Meyer, and Kuratani, 2009; Smith et al., 2013). A more recent study by Smith and Keinath (2015) favors a model of one single WGD in the vertebrate ancestor that was preceded and followed by independent segmental duplications and translocations in the vertebrate ancestor and in the jawed vertebrate lineage. Ambreen, Khalil, and Abbasi (2014) even proposed that at least the *Hox*-bearing chromosomes, a model gene family for the investigation of the 2R-WGD, emerged exclusively from small-scale duplications. It remains unclear, therefore, whether the identified lamprey arrestins represent 1:1 orthologs to the jawed vertebrate arrestins

that resulted from shared segmental and/or WGDs or whether the lamprey arrestins arose from independent duplications after a shared first WGD.

Neo- and subfunctionalization of arrestins during 2R-WGD

The 2R-WGD shaped the molecular machinery of the neuron including many signaling pathways and affected genes that are preferentially expressed in neuronal tissue (Huminiecki and Heldin, 2010). Non-visual arrestins have a pivotal role in some of the pathways remodeled during 2R-WGD: GPCRs (for review see Gurevich et al. (2014)), apoptosis pathway (for review see Kook, Gurevich, and Gurevich (2014)), MAPK (e. g. JNK, ERK, p38; for review, see Strungs and Luttrell (2014)) and a modulatory role in others: JAK/STAT (Sun et al., 2016), Wnt (for review see Schulte, Schambony, and Bryja (2010)), Notch (Mukherjee et al., 2005; Puca et al., 2013), hedgehog (Parathath et al., 2010; Molnar et al., 2011), focal adhesion (Ma et al., 2012; Cleghorn et al., 2015), nuclear hormone receptor (Zhang et al., 2011; Purayil et al., 2015) or the insulin signaling pathway (Luan et al., 2009; Santos-Zas et al., 2013). Obviously, there are plenty of scenarios for functional neo- and subfunctionalization of arrestins. Vertebrate arrestins are one of the most upstream key regulators in signal transduction and possess an extensive interaction network and a broad expression profile. Both are likely the result of a remodeling process as a consequence of the 2R-WGD. Neo- and subfunctionalization of visual arrestins in the phototransduction cascade is especially interesting, as arrestins are in line with several other gene families of the phototransduction cascade, which was expanded by 2R-WGD and thus paved the way for the development of a sophisticated visual system in the vertebrate clade (Larhammar, Nordström, and Larsson, 2009; Lamb et al., 2016). The basic components of the phototransduction cascade have already existed prior to the 2R-WGD with two homologs of the visual opsins (*SWS*, *LW*) as well as one homolog of each of the cyclic nucleotide gated channel (*CNGC*) and phosphodiesterase subunits (*PDE6*), one arrestin, one G protein receptor kinase (*GRK*) and one Gα protein T (*GNAT*) homolog present in the vertebrate ancestor (Lagman et al., 2013; Lamb et al., 2016). The co-expression of arrestins and opsins or the expression of arrestins in photoreceptor cells in several non-vertebrate species such as vase tunicate (Nakagawa et al., 2002; Horie, Orii, and Nakagawa, 2005), different protostome species (insects (Komori et al., 1994; Bentrop et al., 2001), molluscs (Mayeenuddin and Mitchell, 2003; Gomez et al., 2011), crustaceans (Smith et al., 1995)) and the non-bilaterian cnidarian *Hydra magnipapillata* (Plachetzki, Fong, and Oakley, 2012) confirms that arrestin–opsin signaling is evolutionary old and predates the 2R-WGD. A vertebrate novelty in the context of the 2R-WGD is the accommodation of dim and bright light vision. Single-photon resolution in dim light mediated by rods is permitted by the specialization and optimization of components of the phototransduction cascade as well as the rod anatomy towards high efficiency in signal transduction and longevity even under bright light conditions (Korenbrot, 2012; Ingram, Sampath, and Fain, 2016). This is in accordance with the specialized expression of one of the ohnologs of many of the phototransduction gene families in rods after 2R-WGD: *RHO, SAG, GNAT1, PDE6A,B, PDE6G, CNCGCα1, CNCGCβ1, GRK1,* while another ohnolog is preferentially expressed in cones: *OPN1LW, OPN1SWS1, ARR3, GNAT2, PDE6C, PDE6H, CNCGCα3, CNCGCβ3, GRK7* (Nordström, Larsson, and Larhammar, 2004; Larhammar, Nordström, and Larsson, 2009; Lamb et al., 2016). As with lamprey non-visual arrestins, orthology relationships within those gene families are not always clear. Nevertheless, functional information supports the existence of rod and cone-like photoreceptors mediating single-photon resolution in jawless fish such as lampreys (Morshedian

and Fain, 2015). Under the assumption that the rod phototransduction cascade in lampreys is basically similar to the jawed vertebrate rod phototransduction cascade, a loss of the putative *SAG* gene in lampreys as suggested by the paralog absence pattern does not seem likely in the biological context. It is possible that the paralog absence pattern is caused by problems during the sequencing process.
As other phototransduction genes, the visual arrestins arrestin-1 and arrestin-4 were initially proposed to also be specifically expressed in either cones or rods, respectively. Nevertheless, Nikonov et al. (2008) revealed that both visual arrestins are expressed in mouse cones with the concentration of arrestin-1 exceeding arrestin-4 50-fold. The co-expression profile of both visual arrestins in cones and rods of other species is not well studied with partially contradictory results that might be explained by differences in temporo-spatial expression (Nikonov et al., 2008; Amann et al., 2014; Craft et al., 2014). Arrestin-4 is not expressed in rods consistently among all species investigated so far. This spatial subfunctionalization is likely accomplished by the acquisition and loss of regulatory elements such as transcription factor binding sites. Interestingly, apart from genes involved in signaling, transcription factors are known to be preferably retained after 2R-WGD (Huminiecki and Heldin, 2010). As expected, some transcription factor ohnologs that did not drive expression within the eye before 2R-WGD, diverged and gained eye specificity in vertebrates (Holland et al., 2017), e. g. the major regulator of rod photoreceptor gene expression *NRL* (Hao et al., 2012) and *CRX*, which is necessary for photoreceptor differentiation and survival (Corbo et al., 2010). Other transcription factors, such as the transcription repressor *RAX*, a driver of eye morphogenesis, have already been associated with retinal functions in lancelet before 2R-WGD (Orquera and Souza, 2017). Regulatory elements in the promotor region of arrestins have been subject to a few selected studies that identified *RAX*, *CRX*, *NRL* and *VSX2* as regulators of mouse or frog *SAG* expression (Chen et al., 1997; Mani, Besharse, and Knox, 1999; Kimura et al., 2000; Dorval et al., 2006) and several *CRX*-binding elements in conjunction with a TATA element as driver of *ARR3* expression in mouse and human (Zhu et al., 2002; Pickrell et al., 2004). These results are further confirmed and extended by more recent, whole genome-scale Chromatin Immunoprecipitation DNA-Sequencing (ChIP-Seq) studies that revealed binding of *CRX* upstream of all four arrestin genes (Corbo et al., 2010) and binding of the promotor of photoreceptor development, *OTX2*, specifically in proximity to *SAG* (Samuel et al., 2014). Genome-scale ChIP-Seq studies focused on eye transcription factors will highly contribute to a systematic understanding of the expression of eye-specific genes like visual arrestins and the remodeling of eye-specific transcription factor network after 2R-WGD in future. From the information gathered so far, it seems as if *SAG*'s expression is regulated by more transcription factors and thus more strictly controlled in comparison to *ARR3*.
Apart from the retina, both visual arrestins are expressed in the pineal gland (Yamaki et al., 1990; Kroeber, Schomerus, and Korf, 1998; Zhu et al., 2002), an endocrine gland that regulates the circadian rhythm in mammals (Sapède and Cau, 2013). The pineal gland and parapineal organs or parietal eye in fish and reptiles, respectively, are summarized as pineal complex (Solessio and Engbretson, 1993; Lagman et al., 2015). The pineal complex is a vertebrate innovation that evolved from an ancestral vase tunicate-like photoreceptor (Klein, 2006) and expresses a specific set of evolutionary closely related opsins that diversified during 2R-WGD (Hankins, Davies, and Foster, 2014) and are potential interaction partners of arrestins. The monophyletic group of cone- and rod-opsins (visual opsins) emerged from an ancestral *LW* and *SWS* gene that were arranged in tandem before 2R-WGD (Lagman et al., 2013). After 2R-WGD, six ohnologs were retained: *OPN1LW*, *OPN1SW1*, *OPN1SWS2*, *RH1*, *RH2*

and pinopsin. The duplication scenario and orthology relationships of the other three existing opsins (val-opsin, parapinopsin, parietopsin) is less clear as different studies report conflicting tree topologies and usually do not consider a complete set of opsins from single non-vertebrate deuterostome species (Terakita, 2005; Davies, Hankins, and Foster, 2010; Sato et al., 2011). Nevertheless, all studies support the monophyly of those nine opsins with encephalopsin/*OPN3* as closest outgroup.

Interestingly, all vertebrate opsins except for parapinopsin have to be regenerated after absorption of a photon and subsequent isomerization of the chromophore, and cannot be re-activated by absorption of a new photon, i. e. they are monostable or bleaching (Furukawa, Hurley, and Kawamura, 2014). This bleaching ability is unique to vertebrate opsins and differs from invertebrate opsins, which switch between two stable conformations by subsequent photon absorption, i. e. are bistable. The bleaching ability of vertebrate opsins is connected to the reorganization of the interaction interface of opsin amino acids with the protonated chromophore 11-*cis*-retinal, including a shift of the counter ion E113 (invertebrates) to E181 in vertebrates (Terakita, Kawano-Yamashita, and Koyanagi, 2012). To track the shift from non-bleaching to bleaching, Kawano-Yamashita et al. (2011) and Kojima et al. (2017) investigated the bleaching behavior of opsins at key positions within the phylogeny: lampreys and vase tunicate. The vase tunicate opsin, which groups with the vertebrate val-opsin and parapinopsin (Terakita, Kawano-Yamashita, and Koyanagi, 2012), has two counter ions (E113, E181), that work synergistically and cause a behavior intermediate between mono- and bistable opsins (Kojima et al., 2017). The arrestin in the same species, *ARR0*, contains the major CBS and co-localizes with the respective opsin (Horie et al., 2008). Kawano-Yamashita et al. (2011) showed in the lamprey pineal organ that a non-visual arrestin (corresponding to the putative *ARRB2* without the minor CBS) mediates internalization of the bistable parapinopsin, while a visual arrestin (corresponding to the putative *ARR3*) co-localizes with the bleaching rhodopsin and translocates to the outer segment in a light-dependent manner.

Connecting the information about the evolution of opsin bleaching behavior and the CBS conservation pattern of arrestins, I and others postulate a close co-evolution of the bleaching ability of opsins and the clathrin-mediated internalization by arrestins (Kawano-Yamashita et al., 2011; Terakita, Kawano-Yamashita, and Koyanagi, 2012; Kawano-Yamashita, Koyanagi, and Terakita, 2014; Koyanagi et al., 2017). In this Chapter, I provide evolutionary evidence that the major CBS was lost in the ancestor of visual arrestins as hypothesized by Kawano-Yamashita, Koyanagi, and Terakita (2014) and Koyanagi et al. (2017). Acquisition of the bleaching ability in the ancestor of the nine vertebrate opsins thus likely co-occurred with the loss of the major CBS in arrestins. According to this model, parapinopsin might have acquired its bistable nature due to a family-specific reorganization of the interaction network descending from a bleaching ancestor and subsequently gained non-visual arrestin signaling. Further work is necessary to clarify the full opsin repertoire and characterize the interaction networks that determine mono-/bistability in non-vertebrate deuterostome opsins in comparison to non-visual vertebrate opsins.

Bleaching of opsins is a major mechanism for desensitization of vertebrate opsins, that enables (1) Single photon resolution in dim light conditions as no second photon can be absorbed by a bleached opsin; (2) Higher efficiency in G protein activation due to the counter-ion displacement to E181 (Tsukamoto et al., 2009). Both effects contribute to a higher sensitivity necessary for dim light vision in the newly evolved vertebrate rods. The bleaching opsin molecule cannot be re-activated until it re-associates with a 11-*cis*-retinal chromophore that is regenerated from 11-*trans*-retinal in a complex regeneration cycle in the retinal pigment cells (Kiser et al., 2012). For this reason,

internalization of opsins mediated by arrestins, subsequent sorting, recycling and degradation is no longer necessary. Taking this hypothesis further, the bleaching opsins pinopsin, val-opsin and parietopsin that are expressed in the pineal organ like visual arrestins, are potential interaction partners of arrestin-1 and arrestin-4.
I assume that the ancestral visual proto-arrestin possessed an exon–intron structure similar to the vertebrate *ARRB1* and subsequently lost exon 15, which encodes the major CBS. Given that splicing of exon 15 is conserved across vertebrates as shown in this Chapter, why would the loss of this exon in visual arrestins be necessary if the same protein product can be expressed by skipping exon 15 during splicing? Unexpectedly, an answer to this question is provided by a completely different line of research: the study of the permanently active rhodopsin mutant K296E, a naturally occurring mutant causing Retinitis Pigmentosa, a blinding disorder that leads to retinal degeneration (Fahim, Daiger, and Weleber, 1993). Moaven et al. (2013) investigated the interaction of this rhodopsin mutant with arrestin-1 in a mouse model and uncovered that the cell death phenotype can be rescued by expression of the p44 arrestin splice variant lacking the AP-2 binding motif. This study proposed that cytotoxicity is mediated by recruitment of AP-2, a component of the endocytosis machinery, to the arrestin-1/K296E rhodopsin complex. Arrestin-mediated internalization of rhodopsin is also known to cause cell death of fruit fly photoreceptors (Orem and Dolph, 2002; Satoh and Ready, 2005; Kristaponyte et al., 2012). I propose that the loss of the major CBS encoded by exon 15 and a reduction in the affinity for AP-2 binding in the ancestor of visual arrestins was an evolutionary necessity to avoid a permanent recruitment of clathrin and other components of the internalization machinery to wild type (WT) activated rhodopsin in the outer segment. Subsequent internalization and transport of the vertebrate arrestin-(rhod)opsin complex to the photoreceptor cell body and induced cell death are escaped (Moaven et al., 2013). Escape of exon 15 from splicing could result in a small fraction of arrestins that contain a major CBS. As arrestin-1 is one of the most abundant proteins in photoreceptor cells (Song et al., 2011), even a small fraction of arrestins that bind rhodopsin with high affinity could be enough to trigger cell death and have detrimental effects on photoreceptor cells, possibly over a longer time (Moaven et al., 2013).
The minor CBS and AP-2 site conservation patterns revealed in the current study point to the possibility of a low affinity interaction of *SAG* and *ARR3* with clathrin and AP-2. *ARR3* likely has a higher affinity to AP-2 than *SAG*. The functional implication of this potentially higher affinity interaction waits to be elucidated.

3.4.3 Sub- and neofunctionalization as consequence of the 3R-WGD

A third WGD resulted in a further increase in the number of arrestin paralogs in teleosts to six or seven gene copies. The retention rates of arrestins (75 %) in the teleost ancestor is much higher than the retention rate averaged over all genes, that was estimated to be max. 20 % (Glasauer and Neuhauss, 2014; Roux, Liu, and Robinson-Rechavi, 2017). Genes retained after the 3R-WGD are, among others, enriched in the gene ontology term “signaling” (Inoue et al., 2015; Roux, Liu, and Robinson-Rechavi, 2017). Recently, retention was connected to high expression and expression in the nervous system (Roux, Liu, and Robinson-Rechavi, 2017). A high number of interaction partners might also stimulate subfunctionalization and thus ultimately gene retention (Sato, Hashiguchi, and Nishida, 2009). All those apparently retention promoting features apply to arrestins.
Studying the arrestin gene family, I captured one of the few reported examples, where different selection pressures (purifying and positive or purifying and neutral

selection) act on the same gene family in different time windows after the 3R-WGD, namely in the ancestral branches of teleosts, Acanthopterygii and euteleosts. The results of this study illustrate that sub- and neofunctionalization act subsequently or simultaneously to drive innovation in accordance with the neosubfunctionalization model proposed by He and Zhang (2005), Braasch and Postlethwait (2012), and Glasauer and Neuhauss (2014).

All investigated teleost genomes retained four visual arrestins and thus the full set of 3R-WGD ohnologs. The expression data considered supports expression of the ohnolog pairs in different tissues and during different developmental stages, ratifying spatial and temporal subfunctionalization. Laranjeiro and Whitmore (2014) also reported differences in the temporal expression of *SAGa/b* and *ARR3a/b* in zebrafish rods and cones. Spatio-temporal subfunctionalization in embryogenesis is a common process after 3R-WGD that applies to about 87 % of all duplicate genes in zebrafish (Kassahn et al., 2009; Glasauer and Neuhauss, 2014). The spatial subfunctionalization of *ARR3* exceeds the tissue level with *ARR3a* expression in the outer layer of either M- and L-cones and *ARR3b* expression in S- and UV-sensitive cones of zebrafish and carp (Renninger, Gesemann, and Neuhauss, 2011; Tomizuka, Tachibanaki, and Kawamura, 2015). Renninger, Gesemann, and Neuhauss (2011) made a first attempt to functionally characterize *ARR3b* and especially *ARR3a* in zebrafish. Knock-down of *ARR3a* in zebrafish larvae resulted in a prolonged cone response recovery rate and thus a reduced temporal resolution under illumination with visible light, whereas knock-down of *ARR3b* did not show any effect. The authors discuss that this effect could be due to the fact that they primarily capture the dominant photoresponse kinetics of the cone type that expresses *ARR3a* rather than functional differences in photoresponse of both ohnologs. I propose that the expansion and diversification of opsins in teleosts is paralleled by a diversification of expression and function of *ARR3a* and *ARR3b*. This is supported by the expression of both ohnologs in different subsets of cones (Renninger, Gesemann, and Neuhauss, 2011), and my comparative analysis identifying receptor binding and proximal residues to be under positive selection and specificity determining among *ARR3* ohnologs.

The shortening of the C-terminal tail of arrestin-4 represents an interesting change that occurred in the teleost ancestor before 3R-WGD. C-terminally truncated mutants of all four arrestin paralogs are well characterized in literature (section 1.5.2) and represent a preactivated, constitutively active version in comparison to the full-length WT, which binds the GPCR phosphorylation-independent (Han et al., 2001; Kim et al., 2013). The C-terminally truncated salamander arrestin-4 binds phosphorylated, activated rhodopsin and the human M2-muscarinic cholinergic receptor more efficiently (Sutton et al., 2005). Given that this truncation leads to an *ARR3* ortholog of almost exact same length in teleosts, I expect that all teleost *ARR3* also discriminate less efficiently between phosphorylated and non-phosphorylated receptor states. I hypothesize that subfunctionalization following the C-terminal truncation enabled the initial ohnolog retention of *ARR3*. Positive selection on few residues in the ancestral euteleost *ARR3b* might have facilitated a functional change later in evolutionary history, up to 80 my after the 3R-WGD.

In contrast, positive selection acted on a fraction of sites of both *SAG* ohnologs directly after the 3R-WGD illustrating an example of simultaneous sub- and neofunctionalization. Subfunctionalization of both ohnologs on subcellular level was documented by Imanishi, Hisatomi, and Tokunaga (1999) in medaka rods. Furthermore, both ohnologs differ in their temporal expression during the circadian rhythm (Laranjeiro and Whitmore, 2014). Interestingly, expression of *SAGb*, but not *SAGa* is regulated by the transcription factor neurod that putatively binds to the *SAGb*

promotor region (Laranjeiro and Whitmore, 2014) illustrating regulatory differences between both ohnologs. As a second example of functional changes, I find SDPs of phosphate and IP6 binding residues, in agreement with functional studies showing that *SAGa* and *SAGb* have different binding affinities for phosphorylated rhodopsin in carp (Tomizuka, Tachibanaki, and Kawamura, 2015).
In contrast to the visual ohnologs, *ARRB2a/b* show very similar spatial and temporal expression patterns in zebrafish with only minor differences, e. g. in regard to spatial expression in zebrafish primordial germ cells. The nearly identical ohnologs were also shown to have similar functions in modulating the distribution of the chemokine ligand Cxcl12a in zebrafish (Mahabaleshwar et al., 2012). In opposition to zebrafish, *ARRB2* of stickleback and pufferfish carry mutations in key functional motifs presumably impairing their function. Due to sparse sampling in this subbranch, I was not able to test for positive selection or to perform MCA. Those arrestins are candidates for genes that underwent neofunctionalization. Functional studies are necessary to clarify their role *in vivo*.
Sato, Hashiguchi, and Nishida (2009) suggested that this second copy of *ARRB2* in Otomorpha (zebrafish and cave fish, in this work denoted as *ARRB2b*) could have arisen from an independent, local duplication of *ARRB2a* in this clade. Although this scenario is in accordance with my gene tree, I reject the local duplication scenario as (1) The synteny with *cd99l2* located upstream and *Pelp1* located downstream of *ARRB2b* is conserved in all four species (except for stickleback, where *Med11* is located downstream); (2) The local duplication scenario requires an independent duplication event in comparison to the emergence by 3R-WGD, which is most parsimonious. The position within the tree might be caused by long-branch attraction of the highly diverged Percomorphacae *ARRB2b*. The ohnolog pair *ARRB2a,b* was falsely reported as conserved in Otomorpha and Acanthopterygii by another study (La Garcia de Serrana, Mareco, and Johnston, 2014) due to an initial filtering strategy based an ohnolog presence in stickleback and zebrafish. Those discrepancies and inaccuracies in previous studies show the importance and value of a thorough and in-detail annotation of arrestin genes.

3.4.4 Independent gene duplications in different deuterostome orders

Arrestins expanded in deuterostomes not only by large scale segmental or WGDs, but also by tandem duplication (*SAG* in cartilaginous fish, *ARR0* in sea urchins) and retrotransposition (non-visual arrestins in marsupials, see section 1.2.1 for mechanistical details on both processes). As the source of expression data for those species of interest is limited, I can just speculate about putative new functions or a possible subfunctionalization.
ARR0 was duplicated in the ancestor of sea urchins. One of the encoded proteins, sea urchin arrestin-0.1, carries substitutions that probably affect receptor binding, phosphate sensing and, possibly, reduce binding to the clathrin adapter protein AP-2, hinting to a modification of existing functions. The observed tandem duplication seems to be in line with the expansion of arrestin interaction partners in sea urchins. This is exemplified by the overrepresentation of the secretin-like GPCR superfamily (Materna, Berney, and Cameron, 2006) and the rhodopsin-type GPCRs expressed in sensory appendages and the nervous system in purple sea urchin (Raible et al., 2006). So far, nine different opsin genes were identified belonging to seven different subfamilies including an echinoderm-specific opsin-lineage and an r- and c-opsin involved in vision (Delroisse et al., 2014). Apart from GPCRs, regulators of arrestins like the Ras-superfamily of G proteins and the receptor protein tyrosine phosphatases also

underwent lineage-specific duplications hinting at a general expansion of molecules involved in GPCR signaling (Byrum et al., 2006; Fitzpatrick, O'Halloran, and Burnell, 2006). The duplication of known arrestin interaction partners leaves many possibilities for neofunctionalization of *ARR0.1* in different cellular contexts.
Within vertebrates, I revealed a tandem duplication of *SAG* in cartilaginous fish. Most studied cartilaginous fish have a duplex retina, which contains both rods and cones and express rod and cone opsins (Lisney et al., 2012). Adaptations to deep sea conditions are frequent for the respective deep-sea species, e. g. the variation of the cone-to-rod ratio in favor of rods or the shift of wavelength detection towards a shorter wavelength (Davies et al., 2009; Lisney et al., 2012). With no information about the environmental conditions of the ancestral cartilaginous fish, I cannot draw conclusions about the evolutionary advantage acquired by the *SAG* duplication in the context of vision, although it is tempting to speculate that vision in the ancestor of cartilaginous fish has also been rod-dominated with possible specializations of rod cell populations. Interestingly, the 1:2 orthology relationship of ghost shark to human co-occurs for *SAG* and N-acetylmelatonin transferase. Both proteins co-localize with the melatonin-synthesis enzyme N-acetylmelatonin transferase in human pinealocytes, possibly representing the scaffolding molecule for the "melatoninosome" (Maronde et al., 2011). N-acetylmelatonin transferase was duplicated during early vertebrate evolution resulting in two copies: vertebrate and non-vertebrate N-acetylmelatonin transferase. Interestingly, the "non-vertebrate" N-acetylmelatonin-transferase was lost in lamprey and all bony fishes (Falcon et al., 2014), which also have a single *SAG*.
All duplications discussed until now led to an intact, new arrestin gene copy according to the available genomic data. This is different for the two retrogenes in opossum, *ARRB1.2* and *ARRB2.2*, which degraded into pseudogenes. The sequence of the retro-pseudogene *ARRB2.2* is still conserved, ignoring the encoded stop codon. Thus, it could still have a regulatory function if transcribed, e. g. as an anti-sense RNA of the original gene (Johnsson, Morris, and Grandér, 2014). Gene duplication via an mRNA intermediate is mediated by a reverse transcriptase descended from long interspersed elements (LINEs) propagating autonomously within the genome (Kaessmann, Vinckenbosch, and Long (2009), section 1.2.1). The opossum genome is significantly enriched in non-long terminal repeat retrotransposons (29.17 %) in comparison to placental mammals (Mikkelsen et al., 2007) containing about 2,000 known retrocopies (Potrzebowski et al., 2008). Especially L1 retrotransposons, the transposon class mediating retro-gene insertion, comprise a high fraction of retrotransposable elements in opossum with 20 % in comparison to 16.89 % in human (Gentles et al., 2007). The genomes of Tasmanian devil and wallaby have of a similar high fraction of LINEs (33.96 % and 28.6 %, respectively, Nilsson et al. (2012)). The formation of retrogenes from non-visual arrestin parental genes is in accordance with their higher expression in the germline (Pain et al., 2005) as compared to visual arrestins (Storto, 2001; Neuhaus et al., 2006). Interestingly, the *ARRB1.2* retrogene is still intact in wallaby. The similarity of the putative 5′ UTR of the retrogene and the parental *ARRB1.1* gene points to the existence of an upstream open reading frame for *ARRB1* and the insertion of this processed long mRNA isoform into the retrogene locus. The *ARRB1.2* retrogene thus possessed a functional 5′ UTR since its emergence and there was no need to acquire regulatory elements in order to be expressed. This specific feature together with the conservation of the 5′ UTR and the encoded amino acid sequence support the functionality of *ARRB1.2* as a protein-coding gene in wallaby.

3.4.5 Loss of arrestin paralogs in different vertebrate orders

This Chapter establishes the loss/pseudogenization of *ARR3* in the ancestor of afrotherians and xenarthrans supported by synteny information, while the fate of *ARRB2* in birds stays inconclusive. A possible failure in detection of this paralog due to strong sequence divergence in the 50 investigated bird genomes can be excluded as identified sequences and sequence fragments show a high sequence identity to mammalian, amphibian and coelacanth *ARRB2*. This could either point to the degradation of the *ARRB2* gene or to difficulties in sequencing or assembly of this specific region within the bird and lizard genomes. Regions known to cause difficulties in sequencing and assembly are heterochromatin, repeat regions (Treangen and Salzberg, 2012) and GC-rich regions (Botero-Castro et al., 2017). Hoskins et al. (2007) demonstrated the existence of protein-coding genes, ncRNAs and pseudogenes in fruit fly's heterochromatin, which could not be recovered by the initial whole genome sequencing. Beyond that, the most recent update of the fruit fly reference genome again improves annotation of genes in heterochromatic regions and resolves 11 previously fragmented gene annotations (Dos Santos et al., 2014). Botero-Castro et al. (2017) showed very recently that about 15 % of the gene repertoire of birds might be overlooked due to their location in GC-rich regions. The localization of *ARRB2* in such a region is supported by the high sequence conservation of different exons throughout the investigated bird genomes and their localization on extremely short contigs.
The protein Med11, encoded by the putative neighboring gene of *ARRB2*, is part of the mediator complex of RNA polymerase II transcription. As this complex is necessary for transcription in a cell-free system and thus an essential cell component (Zhang et al., 2005), *Med11* is probably also encoded in bird genomes, although it is not detectable. On the other hand, *ARRB2* is not expressed in any of the transcriptome data sets considered in addition to the genomes, which cover different tissues and developmental states of several bird species. This supports a real loss of function as *ARRB2* usually has an ubiquitously high expression. Compensation of *ARRB2*'s function by the highly similar *ARRB1* has been shown in *ARRB2* double-knock-out or knock-down experiments in different contexts, e. g. in signal transduction after opioid receptor activation (Bohn et al., 1999), in lung development (Zhang et al., 2010) or regarding centrosome function (Shankar et al., 2010). Both non-visual arrestins bind numerous GPCRs, clathrin, AP-2, c-Src, JNK3, MKK4 and PSK1 with similar affinities. Their expression overlaps in many tissues and cell types. On the other hand, distinct differences regarding expression level (Gurevich, Benovic, and Gurevich, 2002), specific expression patterns (Gurevich, Benovic, and Gurevich, 2004), subcellular localization (Oakley et al., 2000; Scott et al., 2002), specific non-GPCR interaction partners (Xiao et al., 2007) and the preference for active and phosphorylated receptors (Zhan et al., 2011a) have been revealed confirming some non-redundant functions. The different affinities of arrestin-2 and arrestin-3 to many GPCRs (Oakley et al., 2000) result in their differential desensitization and endocytosis (Kohout et al., 2001; Ahn et al., 2004; Ren et al., 2005; Kuo, Lu, and Fu, 2006). Those joint effects likely cause the often reciprocal outcomes in developmental processes depending on the non-visual arrestin paralog being recruited, e. g. in hematopoiesis (Yue et al., 2009), hedgehog signaling (Chen et al., 2004; Parathath et al., 2010), in the cardiovascular system (see Lymperopoulos and Bathgate (2013), p. 302 ff. and references therein), vascular smooth muscle cell proliferation (Kim et al., 2008) and collagen formation (Lovgren et al., 2011). Collectively, these results suggest, that *ARRB2* mediates specific functions in mammals not being able to be fully replaced by *ARRB1*. Under assumption that *ARRB2* is lost in birds or has an extremely low expression, it seems possible that

ARRB1 might take over some of *ARRB2*'s functions. This hypothesis awaits functional characterization of *ARRB1* in birds to be finally evaluated.
The loss of *ARR3* could be shown explicitly for afrotherians and xenarthrans based on synteny information. Arrestin-4 is specifically expressed in cones and pinealocytes (Craft, Whitmore, and Wiechmann, 1994) as discussed above, where it inactivates phosphorylated cone opsin. Additionally, it interacts with different binding partners near the photoreceptor synapse, e. g. Mdm2, JNK3 (Song, Gurevich, and Gurevich, 2007), calmodulin, microtubule or MKK4, ASK1 (Gurevich et al., 2011) acting as a scaffolding molecule. Arrestin-1 and arrestin-4 are co-expressed in at least some cones in human, primates and mouse rising interest in the investigation of different and redundant functions of both visual arrestins (Craft, 2011; Gurevich et al., 2011). The role of *ARR3* in photoresponse has been characterized in transgenic and mouse knock-out models as well as in knock-down experiments in zebrafish larvae by comparing light response and kinetics of cones and temporal contrast sensitivity in behavioral tests. The response of S-dominant cones of *ARR3* double-knock-out mice to light stimuli is similar to WT mice, while recovery from flashes is greatly slowed down in *SAG*/*ARR3* double-knock-out mice (Nikonov et al., 2008). These and other studies (Brown et al., 2010; Deming et al., 2015a; Deming et al., 2015b) collectively concluded that the opsin desensitization function of arrestin-4 can be fulfilled by arrestin-1 and that at least one visual arrestin is necessary for a normal phototransduction shut-off on the single cell level. Additionally, Shi et al. (2007) showed that arrestin-1 can inactivate S-opsin metaII in transgenic mice expressing S-opsin instead of rhodopsin in rods, although arrestin-1 does not seem to be necessary for dim-flash response in WT cones. Collectively, those studies support the possibility that arrestin-1 could take over arrestin-4's function in afrotherians and xenarthrans if expressed in cones.
A more recent study elucidated that the knock-down of *ARR3a* in zebrafish larvae leads to a greatly prolonged cone response recovery under bright light, resulting in a reduced temporal resolution in behavioral experiments (Renninger, Gesemann, and Neuhauss, 2011). This study was further confirmed by a re-investigation of the phenotype of *ARR3* double-knock-out mice and of *ARR3* double-knock-out mice on a all-cone retina background (*Nrl* double-knock-out) that showed a reduced visual acuity and contrast sensitivity when young (Deming et al., 2015a; Deming et al., 2015b). The same studies elucidated that *ARR3* is necessary for cone long-term survival and that a *ARR3* knock-out causes visual phenotype abnormalities including a reduction in cone number and opsin cone expression. Interestingly, *ARR3* and *SAG* often caused opposite phenotypes on the *Nrl* double-knock-out background in regard to electroretinography amplitudes and opsin survival (Deming et al., 2015b). The authors concluded that the function of *ARR3* and *SAG* differ in regard to non-opsin signaling, where the visual paralogs cannot functionally substitute for each other. Those functions may include vesicle trafficking, the regulation of cone opsin stability/turnover or the developmental and circadian regulation of opsin gene expression (Deming et al., 2015a; Deming et al., 2015b). In fact, some interactions with non-opsin partners, e. g. with Als2Cr4 (Zuniga and Craft, 2010) and Rnd2 (Zuniga and Craft, 2010) are specific to arrestin-4, while the activation of the ATPase NSF is specifically mediated by arrestin-1 resulting in an increase of synaptic vesicle exocytosis at the inner membrane (Huang, Brown, and Craft, 2010). A to-be-discovered role of arrestin-4 in the inner segment of the photoreceptor cells is consistent with its prominent expression at synapses and its less drastic translocation from the inner to the outer segment in bright light conditions as compared to arrestin-1 (Zhu et al., 2002; Zhang et al., 2003). In contrast to arrestin-1, this translocation depends on the presence of guanylate cyclase (Coleman and Semple-Rowland, 2005). While arrestin-1 binding

to phosphorylated rhodopsin is highly specific and selective, arrestin-4 also binds non-opsin GPCRs fairly well *in vitro* (Sutton et al., 2005), which are predominantly located at the synapse. One of those receptors is *DRD4*, which is desensitized by *ARR3*, but not *SAG* upon dopamine stimulation and internalized in conjunction with a non-visual arrestin (Deming et al., 2015b). The evolutionary need for *ARR3* has already been discussed in literature emphasizing further differences between the visual arrestins, namely the ability of *SAG* to self-assemble and the transient binding affinity of *ARR3* to opsins (Gurevich et al., 2011).
The differences pointed out above suggest a rather detrimental effect in case of functional loss of *ARR3* without any functional substitution. Vision in afrotherian and xenartheran species is not well studied. A "model" organism in this clade is elephant, that is active during day and night and possesses a rod-dominated retina with the same set of opsins (*RH1, LWS, SWS1*) as do most placental mammals (Yokoyama et al., 2005; Kuhrt et al., 2017). Kuhrt et al. (2017) showed in an immunohistologic investigation of the elephant retina that all cone opsins co-localize with *SAG* suggesting that *SAG* can substitute *ARR3* in phototransduction shut-off in those cones. As arrestin-4 is structurally and functionally more similar to arrestin-2 than to arrestin-1 (Sutton et al., 2005), I hypothesize that at least some of the non-opsin signaling functions that are fulfilled by *ARR3* in placental mammals, can be taken over by a non-visual arrestin. Eventually, other adaptions like co-evolutionary substitutions in arrestin interaction partners could have evolved in afrotherians and xenarthrans to compensate for the loss of *ARR3*.

3.4.6 Inference of previously unknown interaction partners and isoforms of vertebrate arrestin paralogs

Within this Chapter, I identified several motifs to be conserved in orthology groups, for which the respective function has not been characterized experimentally previously. The first PxxP motif involved in c-Src binding and activation in arrestin-2 is also conserved in the arrestin-3 orthology group. This suggests that both non-visual arrestins bind c-Src in a similar fashion (Luttrell, 1999). As suggested by Strungs and Luttrell (2014), variability of all putative PxxP motifs in arrestin-1 implies that it binds c-Src by a different mechanism.
As stated earlier, *ARR3* and the encoded arrestin-4 represent the least characterized vertebrate arrestin. The conservation of the IP6 binding motif in the arrestin-4 orthology group suggests that it binds IP6 with similar affinity as arrestin-1. Furthermore, the substitution patterns of the AP-2 motif in visual arrestins suggest that arrestin-4 binds AP-2 with higher affinity than arrestin-1, but has a lower affinity to AP-2 than the non-visual arrestins as discussed below.
Moaven et al. (2013) showed that the human and mouse arrestin-1 bind AP-2 with lower affinity than non-visual arrestins due to two substitutions (D374N, R384N) in the otherwise conserved consensus motif [D/E]xxFxxFxxxR. As revealed in this work, the D374N substitution occurred in the ancestor of placental mammals, while all other *SAG* strictly maintain the acidic consensus residue [D/E] at this position. The second residue, R384, is variable in visual arrestins across different clades. I hypothesize that this residue in general contributes to a lower affinity of visual arrestins to AP-2 in comparison to non-visual arrestins. The respective residue contacts AP-2 according to the co-crystal structure of the AP-2 β-appendage and an arrestin peptide (Schmid et al., 2006). Moreover, mutation of the homologous position to alanine in bovine *ARRB2* and fruit fly arrestin-2 affects the sequestration of the β-2 adrenergic receptor (Laporte et al., 2000) and the endocytosis of rhodopsin (Orem and Dolph, 2002),

respectively. The AP-2 motif according to Moaven et al. (2013) is conserved in all *ARR3* apart from a conservative F368[IV] substitution that likely has a subtle effect on AP-2 binding (following a relaxed consensus motif according to Schmid et al. (2006)). Experimental validation is needed to access the affinity of arrestin-4 of different mammalian clades (e. g. rodents, primates, other mammals) to AP-2, as those clades possess slightly different sequence patterns at residues 375–386. It will be interesting to see whether the arrestin-4 AP-2 interaction is relevant *in vivo* given the lower expression of arrestin-4 in cones in comparison to arrestin-1 in rods.
Another motif, which mediates the interaction of arrestin with the endocytosis machinery, is the CBS. Arrestin-4 is expected to bind clathrin with about equal affinity as *ARRB1* without the major CBS as engineered and cloned by Kang et al. (2009). A mutational study in bovine *ARRB1* by Kang et al. (2009) suggests that the L334I substitution in *ARR3* of lobe-finned fish might even increase the binding affinity to clathrin. The same study showed that substitutions observed in mammalian and bird *SAG* (L338F, L342F) decrease the binding affinity to clathrin. Thus, I predict that the binding affinity of *SAG* increases in the following clades (first–lowest): ray-finned fish, birds, mammals, while being overall lower than the affinity of *ARR3* to clathrin.
Besides evaluation of motif conservation, the current work tracks the conservation of different splice variants across deuterostome arrestins. I identified four splice variants that are expressed by different paralogous gene copies and are conserved across almost all investigated paralogs and species. Skipping of exon 13 results in the crystallized and functionally characterized *ARRB1S* isoform. Exon 13 encodes the minor CBS, which might cause differences in the arrangement of molecules on clathrin cages in comparison to the *ARRB1L* isoform (Kang et al., 2009). Conservation of this splice variant hints at a to-be-elucidated role of this exon in visual arrestins. In contrast, exon 15, which encodes the major CBS, cannot be skipped in *ARRB1* and is severely shortened in visual arrestins. As a consequence, all non-visual arrestin isoforms possess the major CBS, while none of the visual arrestins do.

Chapter 4

Improvements on the `ExonMatchSolver`-pipeline

4.1 Motivation

The `ExonMatchSolver`-pipeline (EMS-pipeline) as described in Chapter 2 is a useful tool for the investigation of the evolutionary history of a single gene family, that enables the consideration of gene parts that are situated on different contigs of the same genome. Apart from improvements of arrestin gene annotations in deuterostomes (Chapter 3), the EMS-pipeline has been successfully applied for the re-annotation of the Gα protein family in animals (Lokits et al., 2018). In this study, me and co-workers have investigated the evolution of the *GNA* family, which is encoded by 16 genes in human and is thus much larger than the arrestin family. The *GNA* families' exon–intron structure is conserved within and with few deviations across orthology groups in deuterostomes excluding the *GNAZ* gene and the *GNA12* family, that originated by retrotransposition. The gene family is thus well suited for application of the EMS-pipeline. Based on the re-annotation, my co-workers and I proposed a revisited scenario for the emergence of the five primary *GNA* families. Nucleotide gene trees built based on the updated annotations of the *GNAI* family resolved conflicting gene trees proposed in more coarse-grained and less exhaustive studies (Lagman et al., 2012; Krishnan et al., 2015). Furthermore, the exon focused approach helped tracing exon duplications in the *GNAQ*, *GNA11* and *preGNAI* families, which gave rise to different isoforms with multiple exclusive exons.
The application of the EMS-pipeline in the re-annotation of *GNA* genes did not only demonstrate its advantages, but also revealed several limitations of the current implementation. The *GNAI* and *GNAT* genes originated from a tandem duplication before the vertebrate 2R-whole genome duplication (WGD), where they expanded and were retained with three copies in each group (Lokits et al., 2018). The linkage of the *GNAI* and *GNAT* genes is maintained in all investigated vertebrate genomes (except for lampreys) for the three gene sets, which necessitates the manual processing of these scaffolds in order to apply the EMS-pipeline. Another potential difficulty is the estimation of the paralog number encoded in the respective target genome. By default, the EMS-pipeline expects the same number of paralogs to be encoded in the target genome as are given as input set and encoded in the query genome. The current implementation provides the user-option `-WGD` that can be employed to accommodate an expected duplication of the number of encoded genes in the target genome in comparison to the query as seen as consequence of the 3R-WGDs. Although this option usually works fine e. g. for querying a teleost genome with a tetrapod gene query set, it is currently up to the user to eventually re-run the EMS-pipeline with a lower number of paralogs or to duplicate gene entries in order to accommodate gene losses and segmental or tandem duplications.

In this Chapter, I tackle two limitations of the EMS-pipeline implementation as described in Chapter 2 (referred to as EMS-pipeline Version `1` in the following): (1) Automated estimation of paralog number encoded in the target genome; (2) Automated subdivision of contigs with several encoded gene copies.

4.2 Methods

4.2.1 Estimation of the paralog number

The Paralog-to-Contig-Assignment Problem (PCAP) solved by the EMS-pipeline Version `1` assumes that the number of paralogs is known. Relaxing this assumption leads to an extended form of the PCAP that can informally be described as follows: Given a set of $n_c \in \mathbb{N}$ contigs $\mathcal{C} = \{C_1, \ldots, C_{n_c}\}$, a set of $n_t \in \mathbb{N}$ paralog types $\mathcal{T} = \{T_1, \ldots, T_{n_t}\}$, a set of $n_e \in \mathbb{N}$ exons per paralog type $\mathcal{E} = \{E_1, \ldots, E_{n_e}\}$, an assumed number of paralogs $n_p \geq 1$, and a scoring function $\theta(i, l, k)$, find a mapping of the n_p paralogs onto the n_c contigs and an assignment of the paralogs to a type such that the total score is maximized.
The scoring function $\theta(i, l, k)$ denotes again the bit score of the hit of exon k of paralog type l onto contig i. A complete formal specification of the problem is given below as an Integer Linear Programming (ILP) problem. To this end, three sets of binary variables are considered indexed by contig i, paralog j, and type l:
$P_{ij} = 1$ if and only if paralog j is assigned with contig i, $T_{jl} = 1$ if and only if paralog j is of type l, and $X_{il} = 1$ if and only if contig i contains a paralog of type l. The binary variable Q_{ijk} is introduced additionally, with $Q_{ijk} = 1$ if and only if exon k from paralog j is assigned to contig i. While the variables P_{ij}, T_{jl}, X_{il} represent the associations between contigs, paralogs and paralog types, Q_{ijk} represents the associations between the exons (of a certain paralog) and the contigs.
The assignment is subjected to a series of constraints. First, each contig is associated with one paralog at most, each paralog has to be of a certain type, and each contig can be assigned to only one paralog type at most (Eq. 4.1).

$$\forall i : \sum_{j=1}^{n_p} P_{ij} \leq 1, \quad \forall j : \sum_{l=1}^{n_t} T_{jl} = 1, \quad \forall i : \sum_{l=1}^{n_t} X_{il} \leq 1 \tag{4.1}$$

Second, it has to be assured that a contig i is associated with a paralog j if and only if contig i is associated with some paralog type (Eq. 4.2).

$$\forall i : \sum_{j=1}^{n_p} P_{ij} = \sum_{l=1}^{n_t} X_{il} \tag{4.2}$$

Third, the variables P_{ij}, T_{jl}, and X_{il} have to be linked to assure that in case paralog j is associated with contig i and contig i has an associated paralog of type l, then paralog j must be of type l. Thus, from $P_{ij} = 1$ and $X_{il} = 1$ it follows that $T_{jl} = 1$ (Eq. 4.3).

$$\forall i, j, k : \; P_{ij} + X_{il} - T_{jl} \leq 1 \tag{4.3}$$

Fourth, in order to avoid associating a paralog of type l with a contig that has no exon hit from that paralog type, the following constraint is needed (Eq. 4.4).

$$\forall i, l \text{ s.t. } \nexists k | \theta(i, l, k) > 0 : \; X_{il} = 0 \tag{4.4}$$

Fifth, each exon from a certain paralog is assigned to one contig at most (Eq. 4.5).

$$\forall j,k: \sum_{i=1}^{n_c} Q_{ijk} \leq 1 \tag{4.5}$$

Sixth, if paralog j is associated with contig i, then each exon from paralog j that has a non-zero bit score on that contig i, is assigned to that contig (Eq. 4.6).

$$\forall i,j,k \text{ s.t.} \exists l' | \theta(i,l',k) > 0: \; P_{ij} - Q_{ijk} \leq 0 \tag{4.6}$$

On the other hand, if an exon k from paralog j is associated with contig i, then contig i has to be associated with paralog j (Eq. 4.7).

$$\forall i,j,k: \; Q_{ijk} - P_{ij} \leq 0 \tag{4.7}$$

If for a contig i and an exon k there exists no paralog type that has a non-zero bit score of exon k on contig i, then the following constraint forbids associating exon k (for any paralog j) with contig i. That means if there is no paralog type l' s.t. $\theta(i,l',k) > 0$, then $Q_{i,j,k} = 0$ for any paralog j (Eq. 4.8).

$$\forall i,j,k \text{ s.t. } \nexists l' | \theta(i,l',k) > 0: \; Q_{ijk} = 0 \tag{4.8}$$

Next, the assignment is scored by summing up the similarity scores from all exons of the paralog type that is associated with a certain contig. The unweighted score is calculated as follows (Eq. 4.9).

$$\sum_{i=1}^{n_c} \sum_{l=1}^{n_t} \sum_{k=1}^{n_e} \theta(i,l,k) CT_{il} \tag{4.9}$$

Moreover, the weighting of the corresponding scores of each contig is maintained, which depends on the number of contained exons (Eq. 4.10).

$$\sum_{i=1}^{n_c} \sum_{l=1}^{n_t} \sum_{k=1}^{n_e} \mu_i \theta(i,l,k) X_{il}, \tag{4.10}$$

with $\mu_i = |\{k | \exists l' : \theta(i,l',k) > 0\}|$ being the number of (groups of homologous) exons found on contig i, i.e., those where for at least one paralog type l' the score $\theta(i,l',k) > 0$. In addition to $\theta(i,l,k)$, which favors matches with a high similarity score, the factor μ_i is introduced as previously to prefer assignments with multiple exons found on the same contig (Chapter 2).
A combination of Eq. 4.9 and Eq. 4.10 is used as the objective function of the ILP that is to be maximized over all parameters. Multiplication of the weighted score with the assumed number of paralogs (n_p) assures that the maximum weighted score is the primary criterion of the optimization, which has a high numeric value in comparison to the unweighted score. In case of multiple assignments having the same maximum weighted score, the unweighted score as the secondary criterion results in the selection of the assignment, which has the maximum unweighted score (Eq. 4.11).

$$\max \left(n_p \sum_{i=1}^{n_c} \sum_{l=1}^{n_t} \sum_{k=1}^{n_e} \mu_i \theta(i,l,k) X_{il} + \sum_{i=1}^{n_c} \sum_{l=1}^{n_t} \sum_{k=1}^{n_e} \theta(i,l,k) X_{il} \right) \tag{4.11}$$

Identical optimal solutions in ILP problems can cause significant performance variations (Klotz and Newman, 2013). Therefore, to reduce the number of identical solutions, the associations between paralogs and their corresponding types are constraint. An order for the paralogs and their types is introduced in a way such that if paralog j is associated with type l then all paralogs $j' \leq j$ are associated with a type $l' \leq l$ (Eq. 4.12).

$$\forall j, j', l \text{ s.t.} j' < j : \sum_{l'=1}^{l} T_{j'l'} \geq T_{jl} \tag{4.12}$$

4.2.2 Subdivision of gene loci on the same contig

The identification of gene boundaries (subdivision of gene loci) is a known problem in gene annotation. Many programs have difficulties in identification of tandemly duplicated genes (Chapter 1). In the framework of the EMS-pipeline Version `2`, gene subdivision is considered necessary, if hits of the same paralog- and translated coding exon (TCE)-specific query overlap for at least 5 nt regarding query coverage on the same contig. For identification of compartments, one of the full-length query proteins is queried against the target genome (Fig. 4.1). The coordinates are handed over to the `procompart` tool, which is a component of the employed spliced alignment tool `ProSplign`. If compartments are identified to be located on the same contig, the contig's sequence is divided so that the compartments lie on separate contigs and substitute the original contig's sequence entry in the target genome. In order to account for possible undetected exons, the region between the most downstream hit of the upstream gene and the most upstream hit of the downstream gene is included in both new contig sequences. Due to the consideration of strand-specificity and co-linearity by the spliced alignment tool, the final gene annotation is not influenced by this potential double coverage of sequence.

4.2.3 Implementation details

The EMS-pipeline performs three main steps: (1) Pre-processing; (2) The PCAP implemented as ILP problem (EMS) and (3) Post-processing (Chapter 2). The subdivision of gene loci situated on the same contig is accommodated by querying the target genome with a full-length query paralog during the pre-processing step. In "custom-mode" and "fasta-mode", the resulting `BLAST` output is handed over to the `procompart` tool (Fig. 4.1). In "alignment-mode", the full-length query protein sequences are aligned with `Clustal Omega`, full-length profile Hidden Markov Models (pHMMs) are built with `hmmbuild` and queried against the translated target genome employing `hmmsearch`. As the retrieved coordinates are with respect to the translated target genome, they are converted back to genomic coordinates. The `hmmsearch` output format is reformatted to correspond to the `BLAST` output format, which is handed over to the `procompart` tool. The initial hitlist and the target genome are modified so that those compartments that are identified to be situated on the same contig in the original target genome, are situated on separate contigs in the modified target genome.
The estimation of the target genome paralog number is accommodated by changes within the second step (`ExonMatchSolver`, Fig. 4.1). Instead of running the `ExonMatchSolver` with a fixed number of paralogs i as previously, the `ExonMatchSolver` Version `2` is run n times for all paralog numbers [1, n]. n is specified by the new compulsory parameter `paraMax`, which replaces the user option `WGD`. The parameter establishes an upper bound on the encoded paralog number that

is to be accurately estimated as described in the following. For every run i, the ExonMatchSolver returns several run-specific parameters such as the number of paralogs, mapped contigs, mapped exons, unweighted scores (sum of bit scores of assigned exons) and scores weighted by the number of exons encoded on the same contig. The inclusion of lowly scoring exons for $i > i_{true}$ necessitates a more stringent E-value filtering (E-value < 0.01) in the initial blastall/hmmsearch genome-wide homology search. The EMS-pipeline includes a second, more sensitive homology search step applied to retrieve scores for those exons that scored with certain paralogs on the same contig and are unscored with other paralogs. The E-value of this step is not modified (E-value < 1). The paralog number is estimated from the ExonMatchSolver output of the first iteration, while n is fixed accordingly for the second and third run of the ExonMatchSolver.
I empirically found that the encoded paralog number can be estimated from the unweighted scores in an approach that is conceptually similar to the steepest ascent algorithm. In contrast to steepest ascent, I am given a set of points θ_i of unweighted scores with i being an integer. No continuous function is to be estimated here. Instead, I maximize the gradient (G_i) of all consecutive pairs of points $i-1, i$. The true paralog number i_{true} corresponds to the paralog number that maximizes G_i.

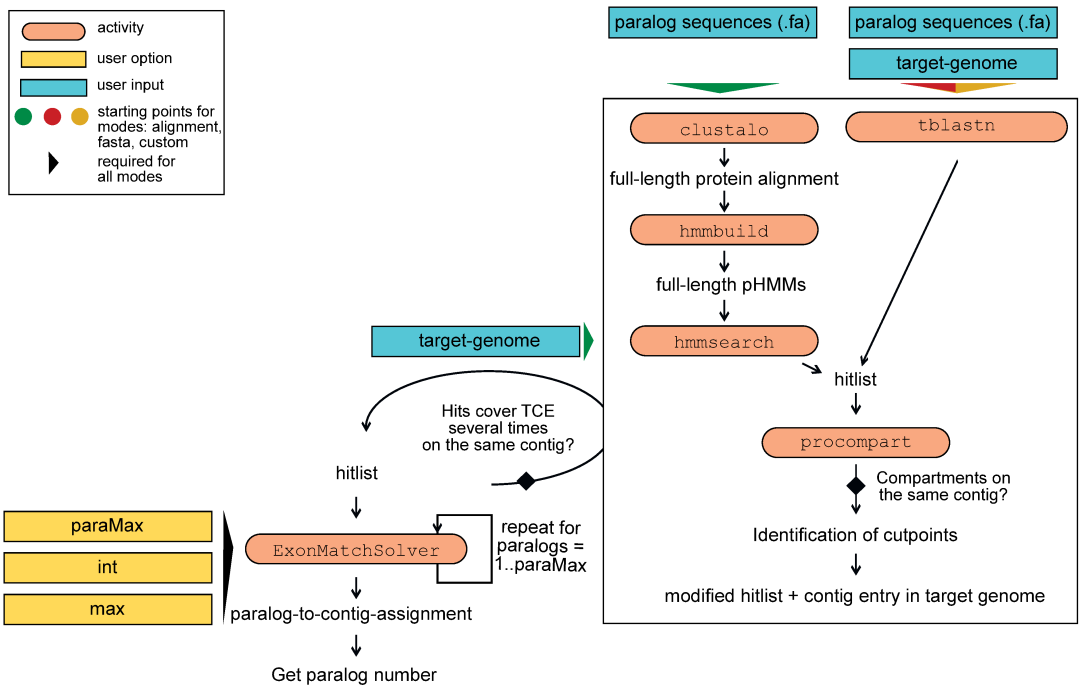

Figure 4.1: Overview about changes of the EMS-pipeline Version 2 with respect to Version 1. The schematic shows only a part of the workflow, specifically the steps "hitlist → ExonMatchSolver → paralog-to-contig-assignment" (see Fig. 2.3 for all steps). The workflow on the right hand side is only entered if hits for the same paralog- and TCE-specific query on the same contig overlap for at least 5 nt regarding the query coverage. User options are given on the left side and above in yellow. Abbreviations: pHMM – hidden Markov Model; TCE – translated coding exon.

4.2.4 Assessment of the `ExonMatchSolver`-pipeline Version 2

The EM-pipeline Version `2` implementation was tested with several real life examples of the arrestin and latrophilin gene families. For those examples, the number of paralogs encoded in the target genomes are known due to manual curation and the orthology relationships can be resolved under consideration of external synteny information given the respective genome annotations. All examples were run in "custom-mode" providing paralog- and TCE-specific sequences and the full-length protein sequences of the protein-coding gene family as input. The new `paraMax` option was set to ten paralogs in all examples except for purple sea urchin (*Strongylocentrotus purpuratus*, `-paraMax 5`). As examples, I annotated arrestins in the orang utan (*Pongo abelii*) and purple sea urchin genomes with the four human arrestins as input (*Homo sapiens*). Furthermore, pufferfish (*Takifugu rubripes*) arrestins and cod (*Gadus morhua*) latrophilins were annotated with the complete and curated gene sets from zebrafish (*Danio rerio*, Chapter 2).

4.3 Results

As shown in Chapter 3, purple sea urchin *ARR0* is duplicated and situated in tandem on *Scaffold_82* (Fig. 4.2 A). The EMS-pipeline Version `1` and the `Scipio` command line Version `1.4.1` recover only a single arrestin (Fig. 4.2 B). The EMS-pipeline Version `1` warns the user that two paralogs might be located on the same scaffold. The subdivision of the scaffold overcomes this problem as it enables the EMS-pipeline Version `2` to retrieve separate gene loci and thus annotations for both genes. The accommodation of the paralog number estimation in Version `2` necessitates a more stringent filtering of the initial hitlist. As expected, this is reflected in the initial `ExonMatchSolver` output, where 12 out of *ARR0.1*'s 15 exons are retrieved in Version `1`, but only eight exons in Version `2`. Nevertheless, the EMS-pipeline Version `2` retrieves eight additional exons during the spliced alignment step and misses only one exon of *ARR0.1* in comparison to Version `1`.
This example is especially difficult in regard to the type assignment as the four human arrestins in the input set have a many:many orthology relationship to the purple sea urchin arrestins. In such cases, the user is encouraged to try different queries with the spliced alignment tool and resolve conflicts during manual assessment of the resulting annotations as different queries in the spliced alignment step might retrieve different results. Specifically, Version `1` queries the locus with *ARRB2*, while Version `2` uses *SAG* as query during the post-processing step. The query with *SAG* causes the EMS-pipeline Version `2` to miss exons 15 and 16 of *ARR0.1*. As exon 15 is present in *ARRB2*, while missing in the human *SAG* gene, the respective exon can only be retrieved with *ARRB2* as query during the spliced alignment step (Fig. 4.2 B, C). Nevertheless, the same query, that missed exon 16 of *ARR0.1*, retrieves exon 16 of *ARR0.2*, an effect likely caused by divergent evolution across the tandem duplicates.
In order to find a good estimator for the paralog number encoded in the target genome, I considered different parameters of the `ExonMatchSolver` solution, namely the number of mapped contigs, of mapped exons, the score weighted by the paralog number and the unweighted score (Eq. 4.9, 4.10), that change in dependence on the given paralog number (Fig. 4.3). The estimation of the encoded arrestin number in the orang utan and pufferfish genomes is comparably easy as arrestins are a self-contained gene family. The applied strict *E*-value cut-off prevents the inclusion of spurious hits into the `ExonMatchSolver` solution. Therefore, all four parameters stagnate at i_{true}, four in the orang utan and seven in the pufferfish example (Fig. 4.3)

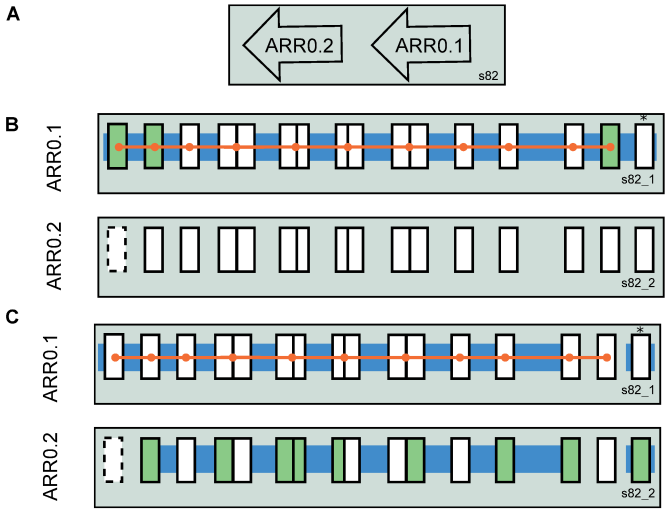

Figure 4.2: Comparison of different versions of the EMS-pipeline predicting arrestin genes in purple sea urchin. Two arrestin genes are encoded on *Scaffold_82* in purple sea urchin (A). The four human arrestins were used as queries for `Scipio`, the EMS-pipeline Version `1` (B) and Version `2` (C). Coding exons of the curated annotation are shown as black open boxes on their respective scaffold (grey boxes). Putatively missing (or deleted) coding exons are denoted by dotted boxes. False positive translated coding exon-hits that were included in the `ExonMatchSolver` solution, but not annotated by the spliced alignment tool are indicated by light green boxes. The solution of the EMS-pipeline considering all of its stages is highlighted by broad blue paths in the back of the exons. `Scipio`'s best scoring proposition is illustrated by colored dots and paths. Note that both gene copies are retrieved with the EMS-pipeline Version `2`, while `Scipio 1.4.1` and the EMS-pipeline Version `1` do not recover *ARR0.2*. Abbreviation: s – scaffold.

In contrast to arrestins, the latrophilin family possesses TCEs with high similarity to TCEs outside of the latrophilin family (Chapter 2). This makes the estimation of the paralog number more difficult. In fact, only the weighted and unweighted score seem to be predictive of the latrophilin number encoded in the cod genome as the number of mapped contigs and exons does not reach a clear plateau at the expected number of six encoded latrophilins. The unweighted score is preferred over the weighted score in order to avoid a bias towards lower paralog numbers in case of high fragmentation with several single exons located on single contigs.
The type assignment based on the paralog-specific bit scores works well in most cases (section 4.2, Tab. 4.1). Consideration of hits retrieved in the more sensitive homology search step as implemented in the pipeline allows for the correct type assignment for orang utan and puffer fish *ARRB1*, which have been falsely assigned to be of type *ARRB2* and *ARRB2b*, respectively in the first `ExonMatchSolver` iteration. In the instance of the highly similar *ARR3* ohnologs of zebrafish, the EMS-pipeline falsely assigns cod *ARR3a* to be of type *ARR3b* (Tab. 4.1).

4.4 Discussion

The EMS-pipeline Version `2` overcomes two limitations that complicated the usage of the previous version. First, a contig's sequence no longer has to be manually divided in order to apply the pipeline to a target genome, where more than one gene family member is situated on the same contig. Secondly, the need to manually

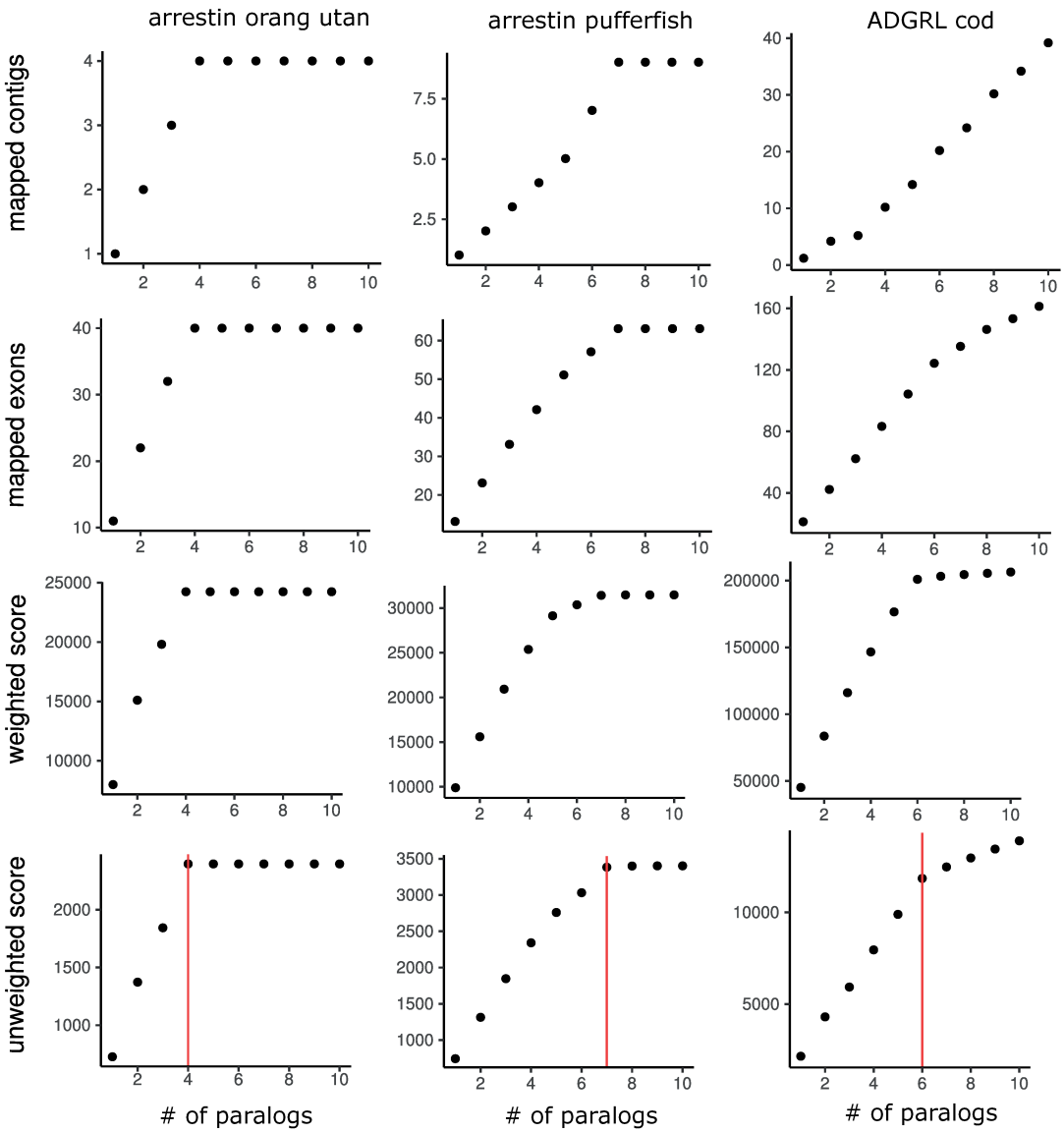

Figure 4.3: Parameters of `ExonMatchSolver` solutions in dependence on the paralog number. Parameter values were collected from the `ExonMatchSolver` solutions after the initial homology search for different fixed numbers of paralogs [1,10] for three different real life examples. The red line marks the paralog number encoded in the target genome according to expert knowledge. The weighted and unweighted scores are good estimators of the paralog number in all examples.

re-run the EMS-pipeline to accommodate unexpected gene duplications and losses is obsolete. Especially gene losses in comparison to the input gene set previously led to the inclusion of spurious hits into the EMS-solution. Estimation of the paralog number encoded in the target genome is accomplished by comparing the unweighted scores of different `ExonMatchSolver` runs with fixed numbers of paralogs. This additional functionality is accompanied by another challenge, the orthology and paralogy assignment. In Version `1`, the input set of proteins was expected to possess 1:1 orthologs in the target genome. In the Version `2` implementation, this 1:1 orthology assignment is no longer possible as the number of genes predicted to be encoded in the target genome and known to be encoded in the query genome might vary. Instead, Version `2` accommodates a "type" assignment based on the score retrieved with

Table 4.1: Type assignments of the ExonMatchSolver (EMS) Version 2 implementation. ExonMatchSolver Version 2 returns a type assignment for every paralog. The type assignment is sometimes incorrect in the initial ExonMatchSolver round, but is usually correct in the final ExonMatchSolver assignment (dark grey). The assignment is incorrect in only one case (light grey) in comparison to the orthology assignment known from consideration of synteny.

Example	Paralog	EMS first iteration	EMS second/third iteration	Correct type
arrestin orang utan	0	*ARR3*	*ARR3*	*ARR3*
	1	*ARRB2*	*ARRB2*	*ARRB2*
	2	*ARRB2*	*ARRB1*	*ARRB1*
	3	*SAG*	*SAG*	*SAG*
arrestin pufferfish	0	*SAGa*	*SAGa*	*SAGa*
	1	*SAGb*	*SAGb*	*SAGb*
	2	*ARRB2a*	*ARRB2a*	*ARRB2a*
	3	*ARRB2b*	*ARRB1*	*ARRB1*
	4	*ARRB2b*	*ARRB2b*	*ARRB2b*
	5	*ARR3a*	*ARR3a*	*ARR3b*
	6	*ARR3a*	*ARR3a*	*ARR3a*
latrophilins cod	0	*ADGRL3a*	*ADGRL3a*	*ADGRL3a*
	1	*ADGRL2a*	*ADGRL2a*	*ADGRL2a*
	2	*ADGRL2b*	*ADGRL2b*	*ADGRL2b*
	3	*ADGRL1a*	*ADGRL1a*	*ADGRL1a*
	4	*ADGRL3b*	*ADGRL3b*	*ADGRL3b*
	5	*ADGRL1b*	*ADGRL1b*	*ADGRL1b*

the different queries, which does not explicitly resolve the orthology and paralogy relationships. Future work will face this problem by re-running EMS-pipeline Version 2 with a set of query and target genomes. This will result in pairwise, directed type assignments between query and target proteins. The corresponding directed graph has k nodes (total number of queries and targets in all considered genomes) that are labeled with the respective scores. The orthology and paralogy relationships can then be directly extracted from the corresponding co-tree. Co-graph editing with a to-be-defined set of rules might be necessary to retrieve a valid co-graph given a known species tree (Hellmuth et al., 2015).

The manual step remaining in the EMS-pipeline workflow is the inspection and comparison of the annotations that are provided by the spliced alignment tools and Scipio in the EMS-pipeline framework. Further improvements could provide a mean to identify and exclude spurious exon hits from the final annotation. A possible direction is the per-exon divergence estimation or consideration of TCE-specific trees across the query protein set or across an orthology group (if considering several target genomes). This process could further simplify the necessary manual inference, which seems unavoidable for high-quality gene annotations and especially in the presence of divergent evolution and 1:many orthology relationships as demonstrated for purple sea urchin arrestins.

Chapter 5

Conclusion and Outlook

Most newly sequenced genomes are not complete and assembled to chromosomes. Genome completeness, assembly and gene annotation quality are limited due to caveats in both, the sequencing process and the assembly of sequencing reads into continuous fragments. In Chapter 2, I have established a tool, the `ExonMatchSolver`-pipeline (EMS-pipeline), that can handle fragmented genomes and assist the assembly of genes distributed across multiple fragments (e. g. contigs). The existence of highly similar genes of the same gene family largely aggravates the annotation and orthology group assignment of genes in fragmented assemblies. The resulting paralog-to-contig assignment problem is NP-hard. The EMS-pipeline accommodates a homology search step with an input gene set consisting of several highly similar paralogs as query. The exon- and paralog-specific hits are associated with a set of homology scores for the target genome. The core of the pipeline (`ExonMatchSolver`) uses an Integer Linear Programming Implementation to solve the paralog-to-contig assignment problem. In short, the objective function of the `ExonMatchSolver` is the maximization of the overall homology scores summed over all contigs, exons and paralogs, whereby the scores are weighted by the number of exons located on the respective contig. Six linear constraints are necessary to describe the biological problem and restrict the solution space.
The EMS-pipeline was successfully applied to simulated data and to two showcase examples in Chapter 2. Especially at high genome fragmentation levels, the tool outperformed a naive assignment method. The run time of the pipeline is in the order of seconds to minutes for biologically relevant exon and paralog numbers. In two biological case studies of the arrestin and latrophilin family, the EMS-pipeline was compared to the `Scipio` pipeline, which also considers fragmentation of genes across different contigs. Nevertheless, `Scipio` did not recover all encoded paralogs in its best solution for neither gene families, but rather proposed the different paralogs to be encoded at the same locus.
The initial implementation of the EMS-pipeline Version `1` was improved in Chapter 4. While the first version with default options predicted exactly as many paralogs in the target genome as given as input paralog set, the EMS-pipeline Version `2` estimates the number of paralogs encoded in the target genome. The estimation is accomplished by running the `ExonMatchSolver` iteratively with different, fixed numbers of paralogs. The overall unweighted scores of the different iterations are then compared and the encoded paralog number is determined automatically. The second new feature concerns handling of several paralogs situated on the same genomic fragment.
In Chapter 3, I have applied the EMS-pipeline Version `1` in a large scale study on the evolution of the arrestin protein family in deuterostomes. Additionally, gene expression data was considered for the determination of arrestin orthology group identity and gene number encoded in the respective genomes. 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. The 2R-whole genome duplication (WGD) in the vertebrate stem lineage gave rise to four arrestin paralogs, which are conserved in almost all clades. Surprisingly, *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 (CBS) and the adapter protein-2 (AP-2) binding motif. The ancestor of the two visual arrestins lost exon 15, which encodes the major CBS. Interestingly, the AP-2 binding site shows deviations from the characterized consensus motif in both visual arrestins. The minor CBS, in contrast, is conserved in only one visual arrestin group (*ARR3*), or shows conservative substitutions.

Although the paralog-to-contig assignment problem is poorly considered during gene annotation, consideration of paralogs that are fragmented across different genomic units and their exon–intron structure is necessary to build high quality gene models. As shown in Chapter 2, `Scipio` has substantial difficulties in distinguishing between close paralogs. A closer inspection of erroneous `Scipio` predictions indicates that these are often the result of incorrect combinations of gene fragments or gene fragments are missing. It therefore seems to be important to explicitly consider and solve the paralog-to-contig assignment problem instead of just selecting best scoring fragments. In particular in the presence of incomplete data and varying sequence divergence across protein sites, simple protein level similarity scores are often insufficient to correctly assign partial or even complete protein sequences to the correct orthology group (paralog type). Treating this issue as an assignment problem as realized in the EMS-pipeline, largely alleviates this particular difficulty of genome annotation. I demonstrated the benefit of the high quality gene models during the investigation of the evolution of arrestins (Chapter 3) and together with co-workers for the Gα protein family (Lokits et al. (2018), not part of this dissertation). Although the evolution of e. g. mammalian arrestins has been examined previously based on database inquiries (Gurevich and Gurevich, 2006a), I uncovered numerous previously unreported gene gain and loss events within arrestins in deuterostomes. Identification of residues that determine specificity 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.

Most available functional protein annotations and mutational studies of a protein family of interest are limited to model organisms such as human, mouse and if the protein family predates deuterostomes, fly. In the era of high-throughput genome sequencing, an evolutionary study of a single protein-coding gene family based on mining of publicly available genomes from multiple species can deliver valuable information about substitutions that were approved during evolution, point to interesting study systems/organisms and broaden the functional understanding of the protein family of interest. The functional understanding is gained by applying different phylogenetic methods to a gene alignment e. g. tree inference, detection of natural selection or detection of specificity determining positions. The effect of missing data on the performance of those tools is mostly unknown, in the worst case it prohibits the tool's usage or causes artifacts. The EMS-pipeline can help in

creating annotations that are more accurate and complete. The tool is applicable to self-contained gene families with a conserved exon–intron structure, which seems to apply to most gene families that emerged during the 2R-WGD.

The insights about the functional evolution of the arrestin family gained include the identification of evolutionary "adjusting screws", positions observed to frequently vary between recent arrestin paralogs and ohnologs and that mediate receptor specificity and influence phosphate binding. Those positions are candidates for the construction of biased arrestins that have already been approved by nature. The deeper understanding of the interactions of arrestins with G protein-coupled receptors (GPCRs) and cytosolic interaction partners and the selective activation of specific downstream pathways is of big interest to pharmaceutical industry. The refined arrestin annotations provided by the work within Chapter 3 are a valuable resource for the increasing arrestin research community (about 3,300 publications with the keyword arrestin in title or abstract as of December 2017). The arrestin alignment should be considered to design arrestin mutants and to select suitable arrestins for experimental studies.

The results and discussion section of Chapter 3 raise three very interesting research questions for arrestin biology pointed out below. Future research should investigate the origin of arrestins. The database analysis conducted in Chapter 3 confirms observations of Mendoza, Sebé-Pedrós, and Ruiz-Trillo (2014), that arrestins predate animals and have already existed in choanoflagellates and Filasterea. As discussed throughout the thesis, a genomic inquiry with dense species sampling in the key species – here pre-animal and early branching animal genomes – is necessary to access the existence of arrestin genes in those species. Although the 2R-WGD led to a spatial subfunctionalization with the visual arrestins being primarily expressed in the retina and pineal gland, the functional connection of opsin, a visual GPCR, and arrestin is evolutionary old and clearly predates bilaterians (Komori et al., 1994; Smith et al., 1995; Bentrop et al., 2001; Nakagawa et al., 2002; Mayeenuddin and Mitchell, 2003; Horie, Orii, and Nakagawa, 2005; Gomez et al., 2011). For example, arrestin and opsin are co-expressed in specialized sensory cells in the non-bilaterian fresh-water polyp (Plachetzki, Fong, and Oakley, 2012). This raises the question whether the ancestral arrestin bound opsin and whether both proteins eventually co-emerged. The earliest opsins have been traced back to animals (Feuda et al., 2012), although Yoshida et al. (2017) recently reported the existence of a rhodopsin in the choanoflagellate *Salpingoeca rosetta* warranting a thorough investigation of the repertoire of opsins in the same genomes as those mined for arrestins. The identified ancestral arrestin could be cloned, expressed and its receptor (opsin) binding behavior experimentally tested.

Another interesting line of future research would concern the functional testing and characterization of the *in vivo* function of the minor CBS and AP-2 binding sites of *ARR3*. It is an open question, whether the minor CBS can mediate low affinity clathrin binding without the major CBS being present or whether this hydrophobic motif might serve as an interaction interface with other proteins.

The acquisition of opsin's bleaching behavior and the loss of the major CBS probably represent co-evolutionary adaptations as discussed in Chapter 3. A detailed characterization of this putative co-evolutionary adaptation further warrants (1) A thorough computational, evolutionary analysis of opsins in non-vertebrate deuterostome genomes to resolve the opsin repertoire in the vertebrate ancestor prior to the 2R-WGD; (2) An experimental characterization of the opsin binding repertoire of *ARR0* in non-vertebrate deuterostomes and (3) An experimental characterization of the binding behavior of visual arrestins to non-visual, bleaching opsins like pinopsin,

parietopsin, val-opsin that are co-expressed in the pineal gland, a brain gland involved in regulation of the circadian rhythm. The rod photoreceptor is an especially well-investigated system in regard to many aspects of arrestin biology, e. g. interaction partners, concentration, translocation and oligomerization (Gurevich et al., 2011), while many questions remain concerning arrestins in the cone photoreceptor and the pineal organ.

For the EMS-pipeline, planned future work concerns the type assignment that arises from the estimation of the number of encoded paralogs in the target genome as discussed in Chapter 4. Further improvements of the EMS-pipeline could include a variant of the spliced alignment step. Instead of performing a homology search again, the available `BLAST` or `hmmsearch` hits could be used directly. The recently developed `GeMoMa` pipeline pursues such an approach (Keilwagen et al., 2016). Comparably to the EMS-pipeline, `GeMoMa` takes advantage of the conservation of intron positions of homologous genes to retrieve gene predictions that are more accurate than predictions from other tools, e. g. `exonerate` or `genBlastG` (Keilwagen et al., 2016). The tool decomposes the query gene into coding exons (equivalent to translated coding exons in this work) that are blasted against the target genome. Furthermore, the two approaches, `GeMoMa` and the EMS-pipeline, could complement each other. While the EMS-pipeline considers gene fragmentation and orthology prediction, `GeMoMa` offers a more accurate approach for computation of spliced alignments. A combination of both tools into a versatile all-in-one tool seems desirable.

This dissertation has drawn attention to a mostly unregarded problem in gene annotation: fragmentation of genes and distribution to multiple contigs. With the EMS-pipeline, I have established the first tool, that explicitly addresses this problem for gene families with highly similar members. The strength of the tool has been exemplified when applied to the annotation of the arrestin family, but waits to be applied to other gene families of special interest.

Appendix A

Additional figures

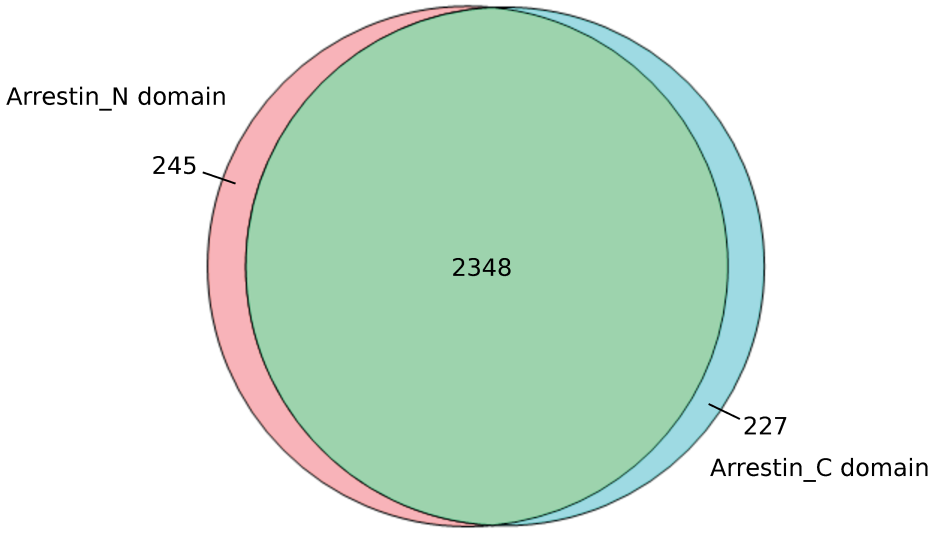

Figure A.1: Scan of the `UniProtKB` with full-length arrestin profile Hidden Markov Models employing `jackhmmer`. `Jackhmmer` results after three iterations were filtered according to parameters specified in section 3.2.1 resulting in 2962 hits in total. Those hits show a good overlap with the `Pfam` *arrestin_N* and *arrestin_C* domains (intersection, light green). The hits that did not contain any of the two domains (142) were excluded for estimation of paralog numbers and are not shown.

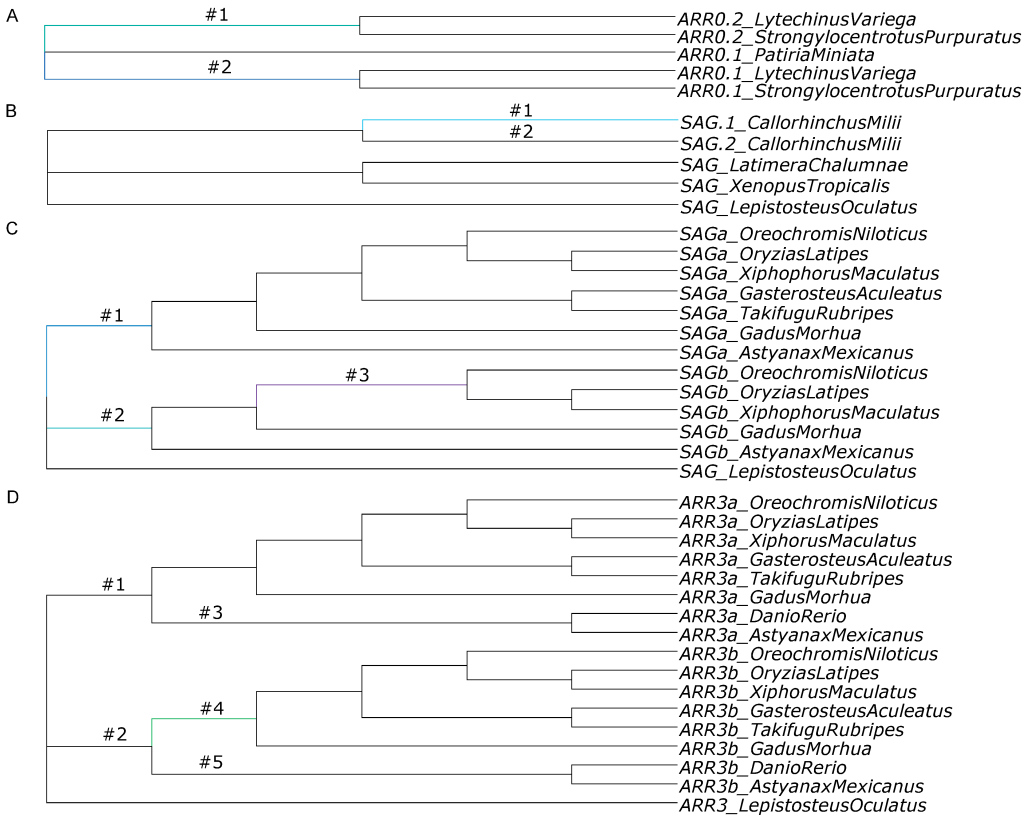

Figure A.2: Foreground branches in the natural selection analysis of arrestins. Specific branches within the arrestin gene trees were tested for positive selection using the branch-site model of `codeml`, part of the `PAML` program. Foreground branches in the separate analysis are marked by numbers. One analysis was conducted per foreground branch with all other branches of the tree in the background. Branches, which were identified to be under positive selection, are colored (Tab. B.6).

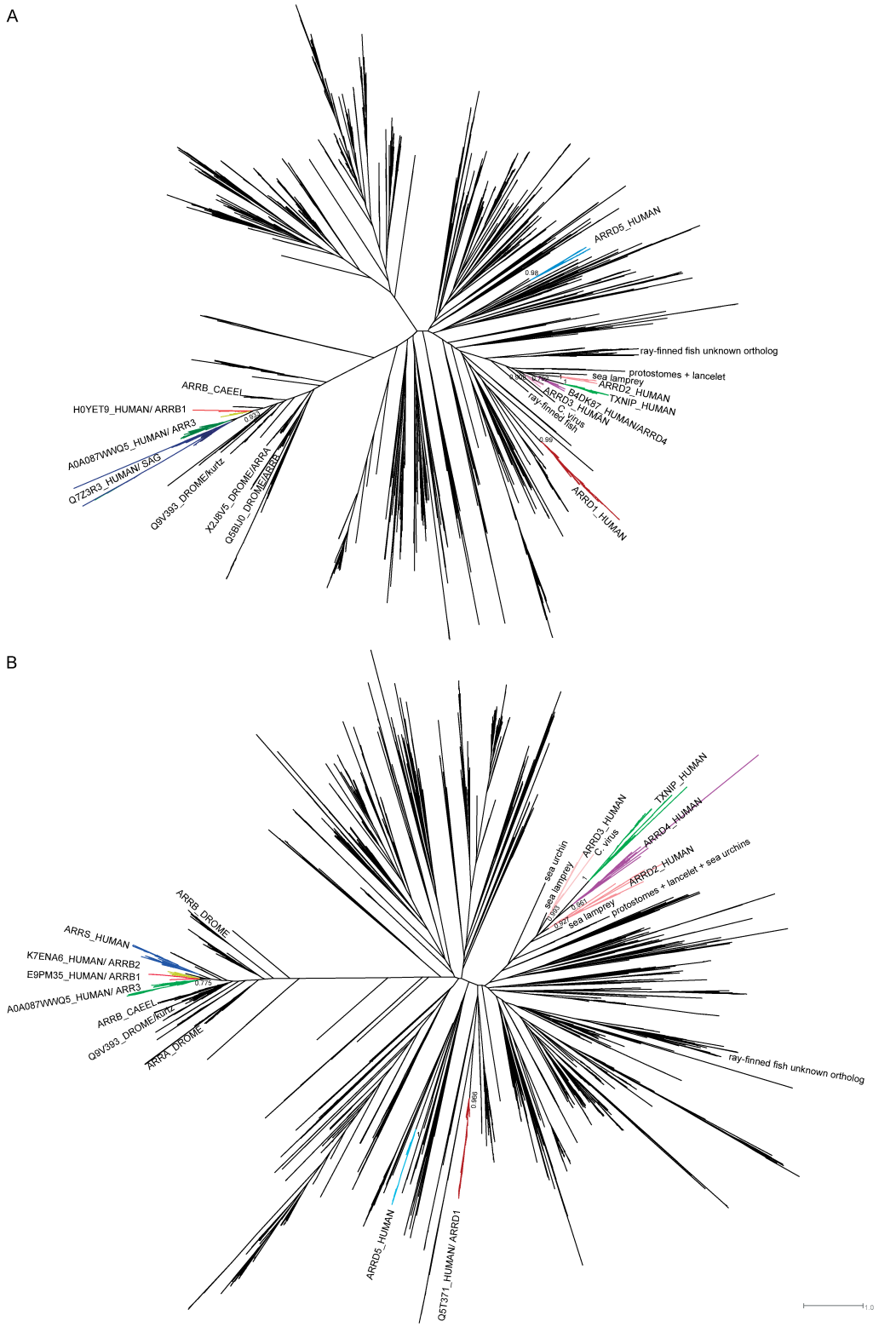

Figure A.3: Approximate maximum likelihood tree of the single arrestin domain hits. A – *Arrestin_C* domain, B – *Arrestin_N* domain (section 3.2.4). Bootstrap support values are shown for splits that contain groups with vertebrate arrestin fold members. The groups with human arrestin fold protein members and their respective 1:1 orthologs are marked in color. Those groups are monophyletic with the following exceptions: *arrestin_C* domain of *SAG* contains 22 nematode sequences, *arrestin_N* domain of *ARRD4* contains one beetle sequence, *arrestin_N* domain contains the *C. virus* sequence. Abbreviations: CAEEL – round worm (*Caenorhabditis elegans*); DROME – fruit fly (*Drosophila melanogaster*).

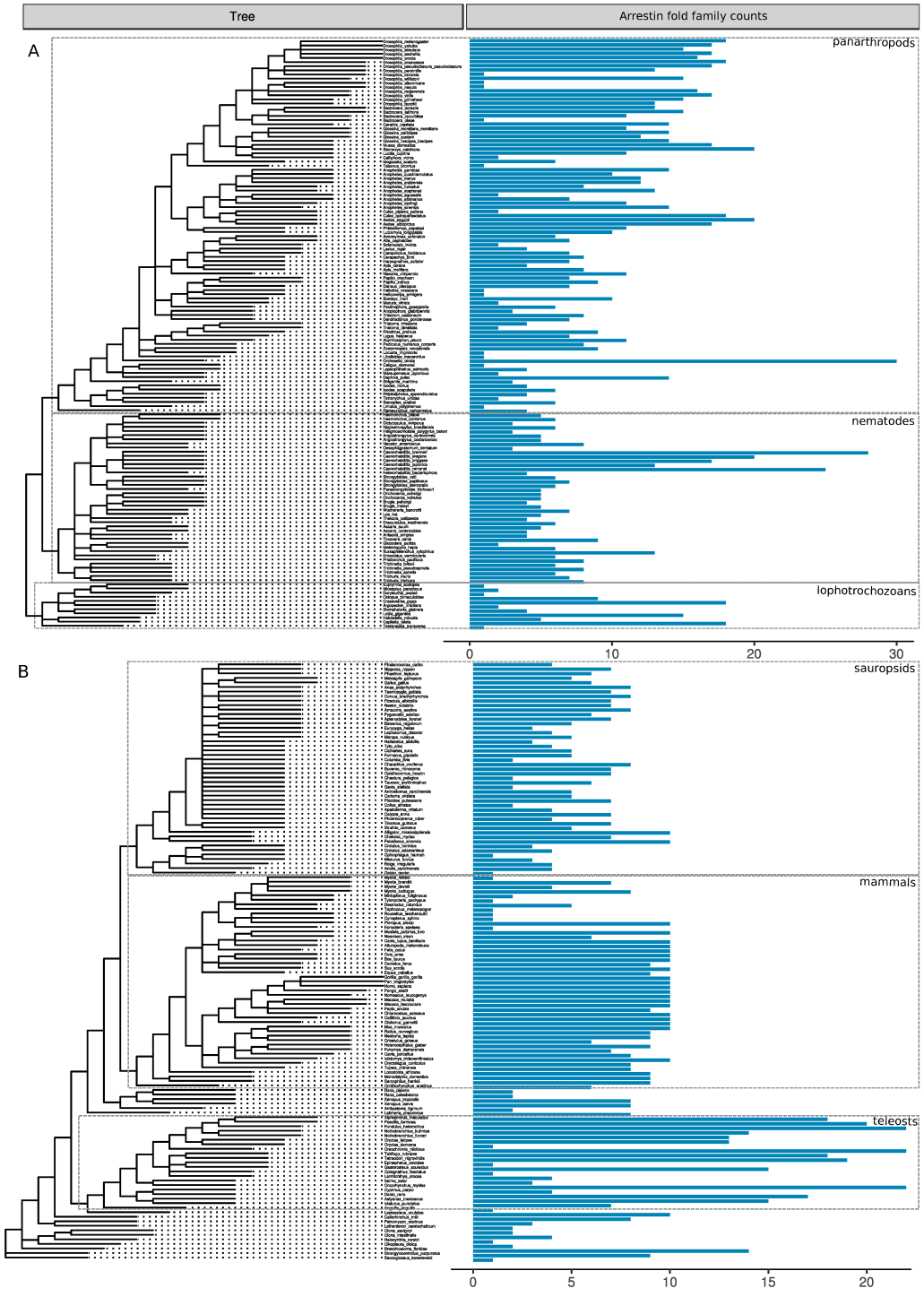

Figure A.4: Abundance of arrestin fold family members in bilaterians according to `UniProtKB`. Hits were assigned to the arrestin fold family if they contained either an *arrestin_N* or an *arrestin_C* domain (section 3.2.1) and were mapped to the `NCBI` taxonomy of protostomes (A) and deuterostomes (B).

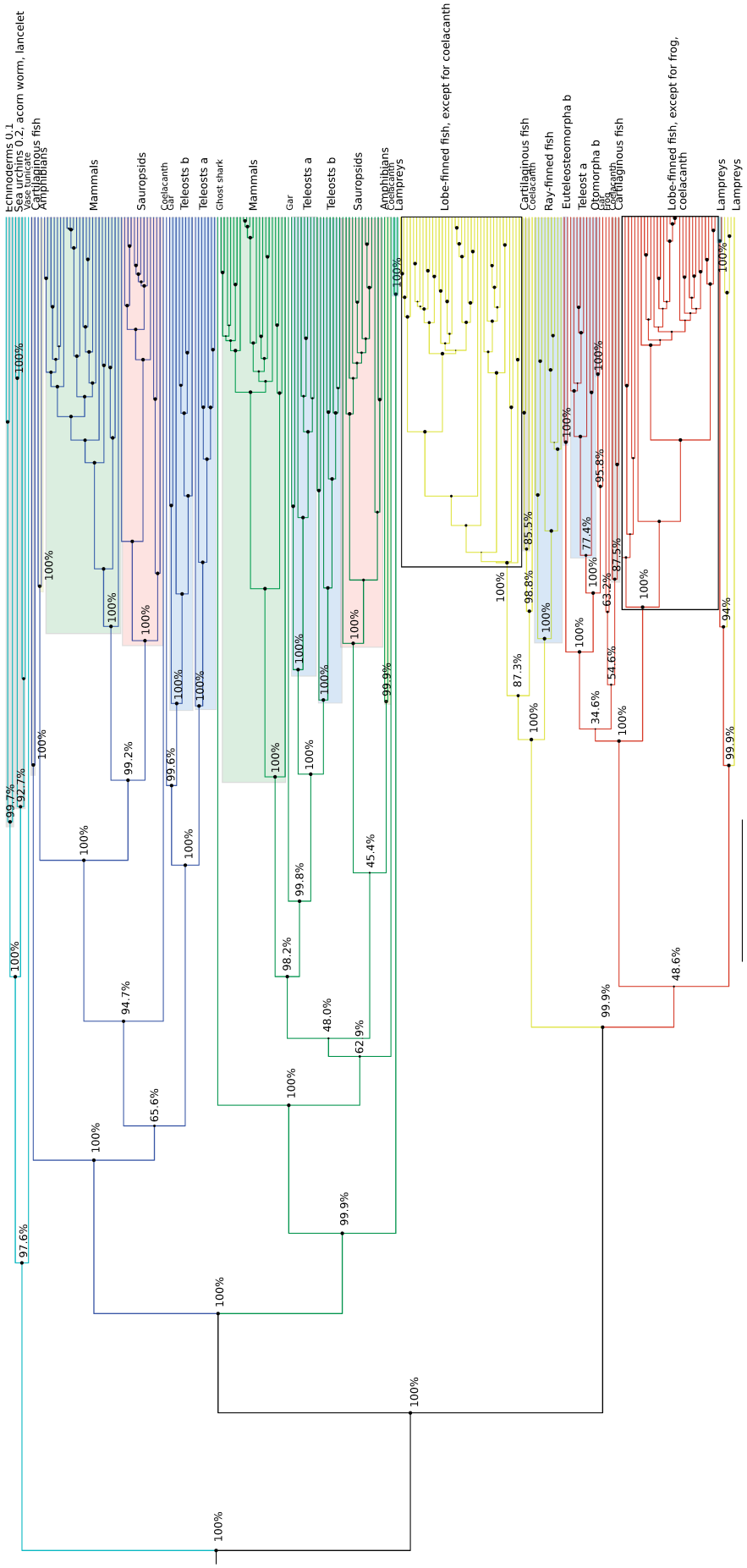

Figure A.5: Rooted bayesian tree of arrestins. The tree was constructed from a codon-based nucleotide alignment of deuterostome arrestins using BEAST (Birth-death model with relaxed log normal molecular clock, substitution model: HKY, estimated shape and substitution rates for each codon partition). Furthermore, vertebrate arrestins were constrained to be monophyletic. The branches of the five main orthology groups are drawn in different colors: *ARR0* – turquoise, *SAG* – blue, *ARRB1* – yellow, *ARRB2* – red, *ARR3* – green. All highlighted species groups are monophyletic except for some *ARR0*, that are summarized for better readability. The size of the dot at all nodes scales to the respective posterior probability.

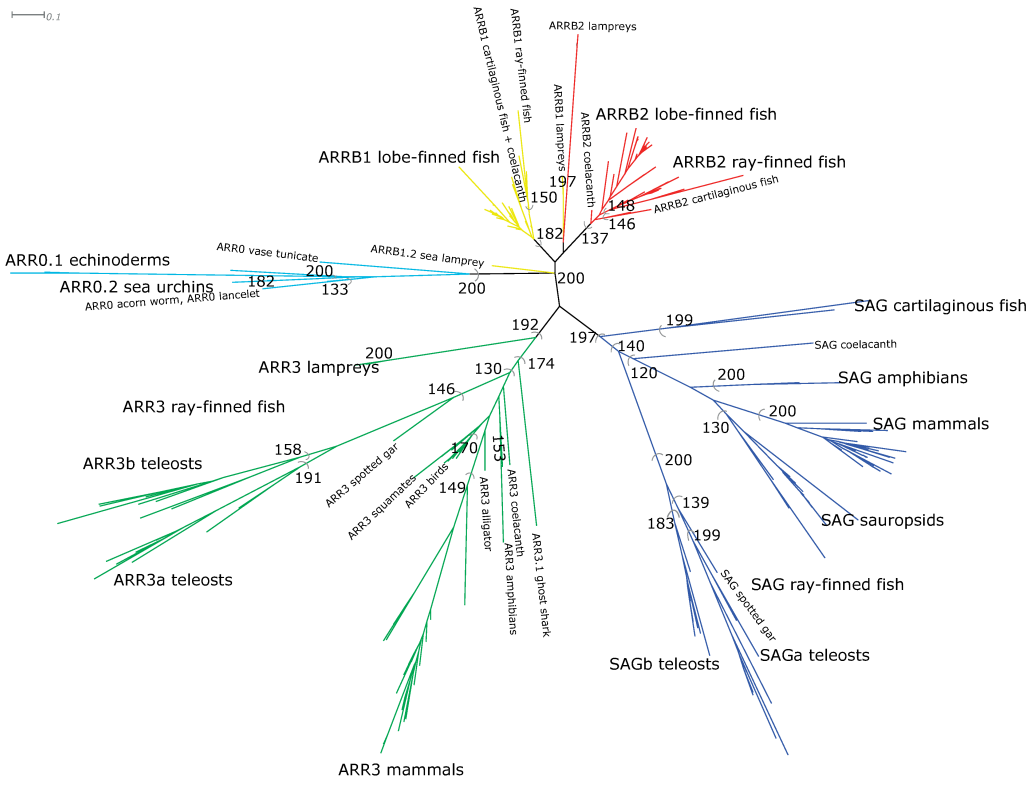

Figure A.6: Maximum likelihood tree of partial arrestins under exclusion of receptor specificity columns. The tree was constructed from an amino acid alignment of deuterostome arrestins using `PhyML` (model JTT+I+G with α 1.04, 5% of invariable sites and 200 bootstraps). All columns that are known to confer receptor specificity were excluded prior to the alignment (80 columns were removed). The different monophyletic and well supported orthology groups are highlighted in different colors. Bootstrap support values from 50...100% are shown for the labeled monophyletic groups. Although bootstrap support values are generally lower than for the full-length arrestin alignment, major splits of the vertebrate animal species tree are still well resolved.

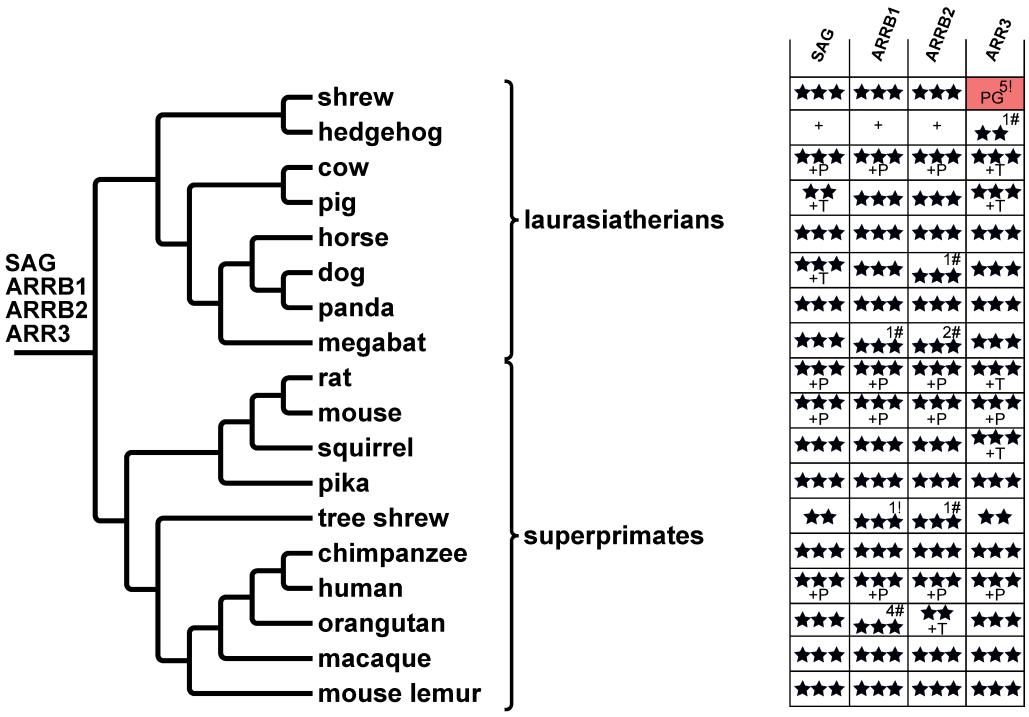

Figure A.7: Arrestin paralogs within laurasiatherians and superprimates. Four arrestin paralogs are encoded in the genomes of both mammalian clades with one exception, *ARR3* in shrew. The gene is probably degraded to a pseudogene (red box). It is not clear, whether this is also true for hedgehog, which has a highly fragmented *ARR3* locus likely due to missing data. The table on the right side of the figure depicts the completeness of arrestin annotations in the respective genomes. Additional support for arrestins from reviewed entries of `UniProtKB` are given in the table. See caption Fig. 3.7/3.10 for an additional description of symbols. The phylogenetic tree was created with `Treegraph 2.0.54` (Stover and Muller, 2010).

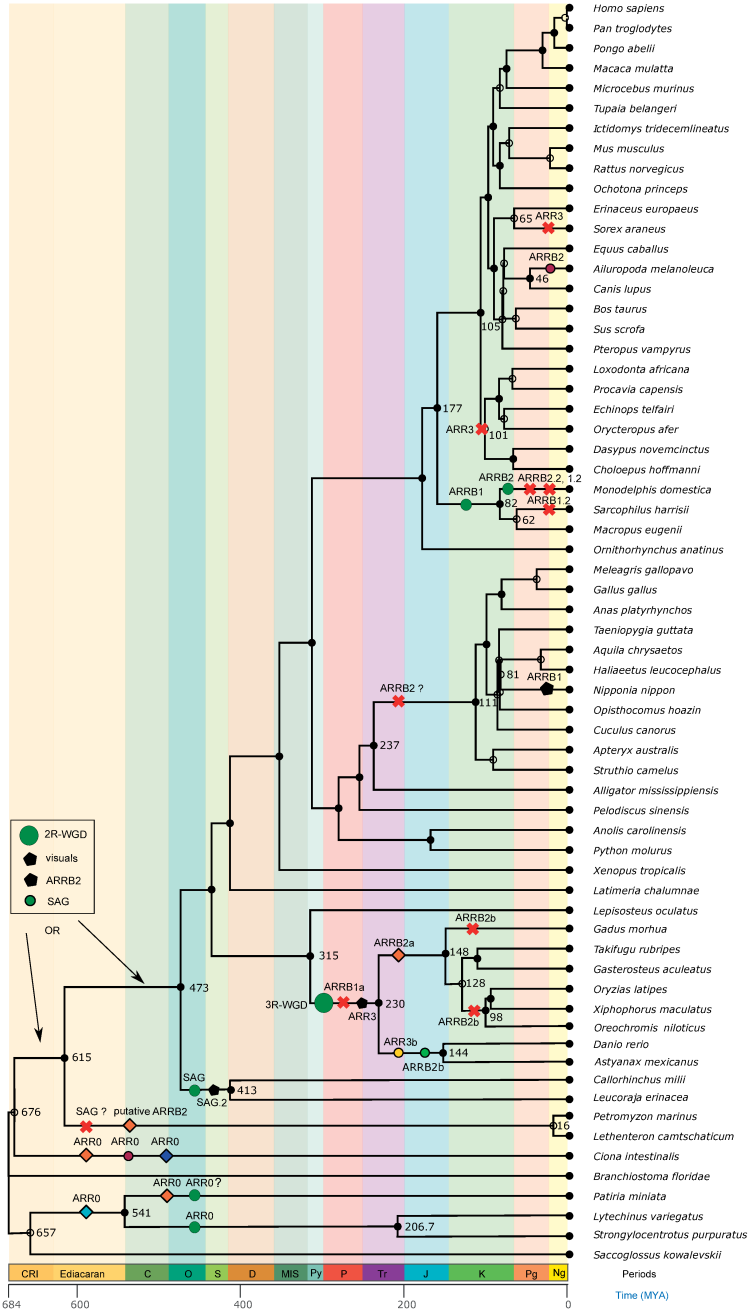

Figure A.8: Summary of arrestin gene, exon and intron gain and loss events in deuterostomes. All events based on the updated annotation and discussed throughout Chapter 3 were mapped onto a timed species tree (Kumar et al., 2017). Trifurcations are labeled with the estimated speciation time, whenever events happened on nearby branches. Crosses and dark green dots indicate gene duplication and loss events, while colored diamonds, dots and pentagons on tree branches symbolize intron gain, intron loss and exon loss events, respectively. Please see Fig. 3.16 for details on changes of the gene structure as well as exon color code. Note that events on the specific branches are placed arbitrarily in regard to the order and exact timing. The sea lamprey-specific non-visual arrestin is omitted from consideration. The figure was created using the timetree webserver. Abbreviations: MYA – million years ago; WGD – whole genome duplication; CRI – Cryogenian; C – Cambrian; O – Ordovician; S – Silurian; D – Devonian; M – Mississipian; Ps – Pennsylvanian; P – Permian; Tr – Triassic; J – Jurassic; K – Cretaceous; Pg – Paleogene; Ng – Neogene.

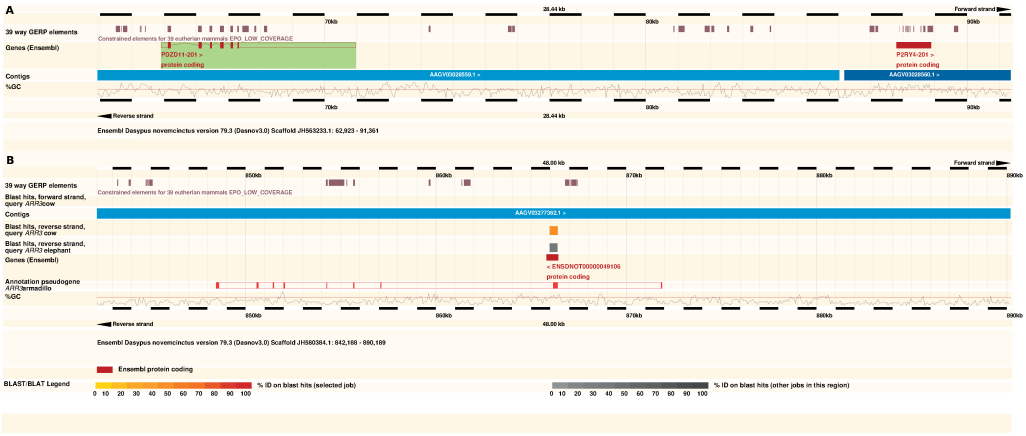

Figure A.9: Candidate loci and genes for *ARR3* in armadillo. A – The genomic region between *PDZD11* (green box) and *P2RY4* of armadillo has no similarity to bovine *ARR3*. B – Instead, bovine *ARR3* returned a blast hit on `JH580384.1` overlapping the `Ensembl` gene *ENSDNOT00000049106* that has an *arrestin_N* domain (PF00339). Annotation attempts using `ProSplign` resulted in annotation of a hypothetical pseudogene (bright red) that contains three internal stop codons and three frame shifts. The picture was generated with the `Ensembl genome browser`. Abbreviations: GERP elements – constrained, conserved elements called by `Ensembl`; ID – identity; kb – kilobases.

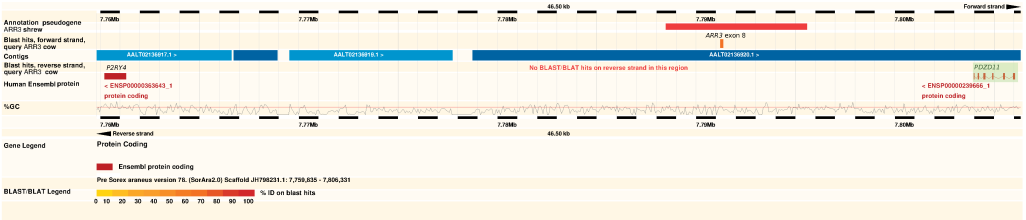

Figure A.10: Genomic locus of the *ARR3* pseudogene in shrew. The respective locus next to *PDZD11* was identified by `tblastn` of bovine *ARR3* against the shrew genome (green box). Homology search using *ARR3* from dog, human and mouse revealed fragments similar to exons 3, 8, 10, 12 and 14. Attempts to annotate the full coding sequence with `ProSplign` resulted in an annotation with at least five internal stop codons. The region spanning the putative exons 3-14 is therefore proposed to represent an *ARR3* pseudogene (bright red). Note that the contig is ungaped in this region. The picture was generated with the `Pre!Ensembl genome browser`. Abbreviations: ID – identity; kb – kilobases.

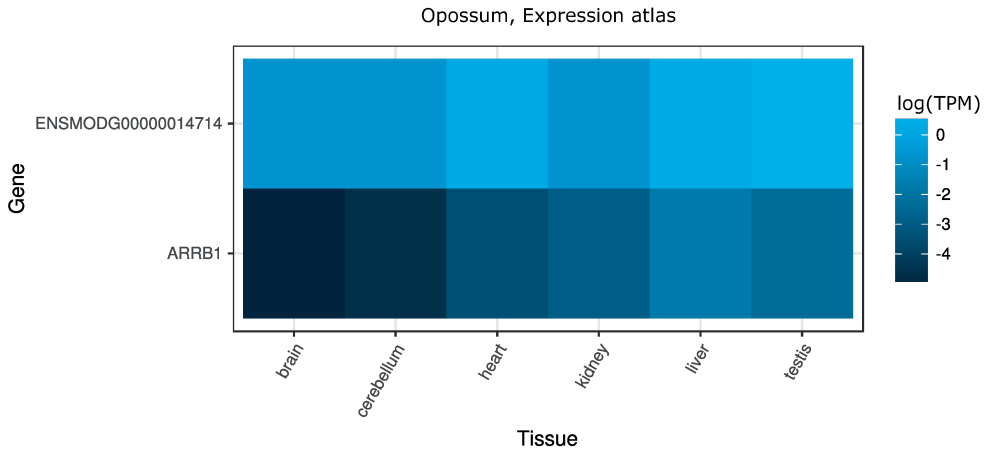

Figure A.11: Expression pattern of *SAG* and *ARRB1* in opossum. Expression values with TPM > 0.5 are shown for different organs as extracted from the Expression atlas (Petryszak et al., 2016). *ARR3* is not expressed above this threshold. As expected from other mammals (human, mouse), the non-visual *ARRB1* has a much higher expression in the shown organs than the visual *SAG* (*ENSMOD00000014714*).

Figure A.12: Structure and genomic locus of the *ARRB1.2* retrogene in wallaby. The first row shows the genomic position of the coding sequence of the *ARRB1.2* gene as proposed in this study (bright red). It is in good accordance with arrestin sequences from `UniProtKB` (yellow), vertebrate cDNAs from `ENA` (green), and expressed sequence tag (EST) clusters from `Unigene` (green). In addition a fraction of cDNAs and EST clusters also show high sequence conservation to the region upstream of the proposed coding sequence. The "Gene (`Ensembl`)" track (dark red) denotes the protein-coding gene *ARRB1-201*. Light pink boxes show high scoring `BLASTz` hits of the opossum arrestin loci against the *ARRB1.2* locus in wallaby. Notice that *ARRB1.2* of opossum likely is a retrogene sharing the retrotransposition event with wallaby. The figure was generated with the `Ensembl genome browser`. Abbreviations: ID – identity; kb – kilobases.

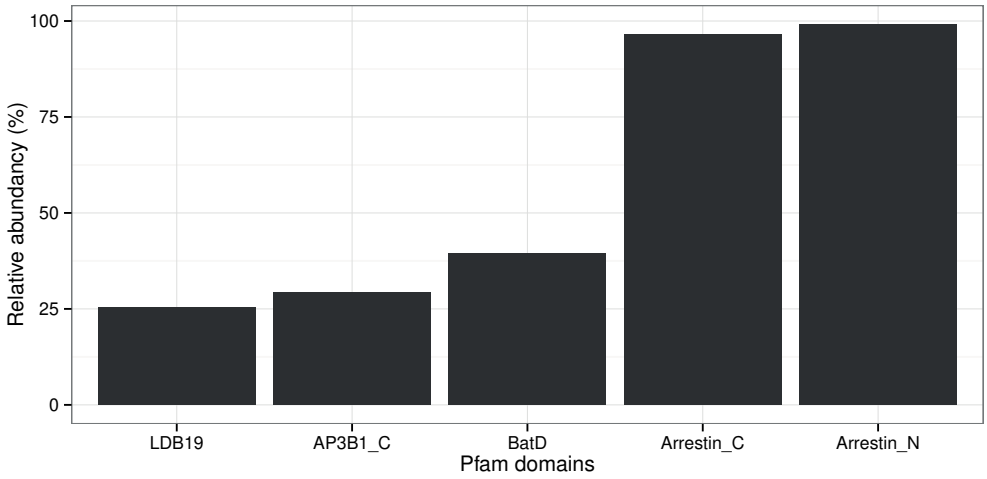

Figure A.13: `Pfam` domains in deuterostome arrestins. All deuterostome arrestins (excluding pseudogenes and fragments of *ARRB2* in birds) were scanned against the `Pfam 28.0` database (Finn et al., 2014). The relative abundance of domains, present in at least 25% of all deuterostome arrestins, is shown.

Appendix B

Additional tables

Table B.1: List of genomes considered for a refined annotation of arrestins. Latin and trivial names are provided together with the version used and source of investigated assemblies. The three or four letter abbreviation in parenthesis are used throughout the document. Additionally, all other 39 genomes from the Avian genomics project were investigated.

Latin name	Trivial name	Genome version	Genome source
Ailuropoda melanoleuca (Aime)	panda	*ailMel1*	Ensembl
Alligator mississippiensis (Ami)	alligator	*v0.1d27*	ftp://ftp.crocgenomes.org/pub/ICGWG/Genome_drafts/alligator.old/amiss_v0.1d27/amiss_v0.1d27.fa
Anas platyrhynchos (Apl)	duck	*BGI_duck_1.0*	Ensembl
Anolis carolinensis (Aca)	anole lizard	*AnoCar2.0*	Ensembl
Apteryx australis mantelli (Aau)	kiwi	*v1.0*	NCBI/GCF_001039765.1
Aquila chrysaetos (Ach)	golden eagle	*v1.0.2*	NCBI/GCF_000766835.1
Astyanax mexicanus (Asme)	cave fish	*AstMex102*	Ensembl
Bos taurus (Bta)	cow	*UMD3.1*	Ensembl
Branchiostoma floridae (Bfl)	lancelet	*Brafl1_v2.0*	http://genome.jgi-psf.org/Brafl1/Brafl1.download.html Branchiostoma_floridae_v2.0.assembly.fasta.gz
Callorhinchus milii (Cmi)	ghost shark	*calMil1.fa*	UCSC (calMil1.fa.gz)
Canis familiaris (Cfa)	dog	*CanFam3.1*	Ensembl
Choloepus hoffmanni (Cho)	sloth	*choHof1*	Ensembl

Continued on next page

Table B.1 – continued from previous page

Latin name	Trivial name	Genome version	Genome source
Ciona intestinalis (Cin)	vase tunicate	*JGI2*	Ensembl
Cuculus canorus (Ccu)	cuckoo	*v1.0*	http://avian.genomics.cn/en/jsp/database.shtml
Danio rerio (Dre)	zebrafish	*Zv9*	Ensembl
Dasypus novemcinctus (Dno)	armadillo	*Dasnov3.0*	Ensembl
Echinops telfairi (Ete)	tenrec	*TENREC*	Ensembl
Equus caballus (Eca)	horse	*Equ Cab 2*	Ensembl
Erinaceus europaeus (Eeu)	hedgehog	*eriEur1*	Ensembl
Gadus morhua (Gmo)	cod	*gadMor1*	Ensembl
Gallus gallus (Gga)	chicken	*Galgal4*	Ensembl
Gasterosteus aculeatus (Gac)	stickleback	*BROADS1*	Ensembl
Haliaeetus leucocephalus (Hle)	bald eagle	*v1.0*	http://avian.genomics.cn/en/jsp/database.shtml
Homo sapiens (Hsa)	human	*GRCh38*	Ensembl
Ictidomys tridecemlineatus (Itr)	squirrel	*spetri2*	Ensembl
Latimeria chalumnae (Lch)	coelacanth	*LatCha1*	Ensembl
Leucoraja erinacea (Ler)	little skate	*v1.0*	NCBI/GCA_000238235.1
Lepisosteus oculatus (Loc)	spotted gar	*LepOcu1*	Ensembl
Lethenteron camtschaticum (Lca)	arctic lamprey	*v1.0*	NCBI/GCA_000466285.1
Loxodonta africana (Laf)	elephant	*Loxafr3.0*	Ensembl
Lytechinus variegatus (Lva)	green sea urchin	*v0.4*	http://www.echinobase.org/Echinobase/LvDownload (Lvar_0.4.20110428.linear.fa)
Macaca mulatta (Mamu)	macaque	*MMUL 1.0*	Ensembl
Macropus eugenii (Meu)	wallaby	*Meug_1.0*	Ensembl
Meleagris gallopavo (Mga)	turkey	*Turkey_2.01*	Ensembl

Continued on next page

Table B.1 – continued from previous page

Latin name	Trivial name	Genome version	Genome source
Meleagris gallopavo (Mga)	turkey	*Turkey_5.0*	NCBI/GCF_000146615.2
Microcebus murinus (Mimu)	mouse lemur	*micMur1*	Ensembl
Monodelphis domestica (Mdo)	opossum	*monDom5*	Ensembl
Mus musculus (Mumu)	mouse	*GRCm38.p1*	Ensembl
Nipponia nippon (Nni)	ibis	*v1.0*	http://avian.genomics.cn/en/jsp/database.shtml
Ochotona princeps (Opr)	pika	*OchPri3*	Pre!Ensembl/ GCA_000292845
Opisthocomus hoazin (Oho)	hoatzin	*v1.0*	http://avian.genomics.cn/en/jsp/database.shtml
Oreochromis niloticus (Oni)	tilapia	*Orenil1.0*	Ensembl
Ornithorhynchus anatinus (Oan)	platypus	*OANA5*	Ensembl
Orycteropus afer afer (Oaf)	aardvark	*OryAfe1.0*	Pre!Ensembl/ GCA_000298275.1
Oryzias latipes (Ola)	medaka	*MEDAKA1*	Ensembl
Pan troglodytes (Ptr)	chimpanzee	*CHIMP2.1.4*	Ensembl
Patiria miniata (Pmi)	bat star	*v1.0*	http://www.echinobase.org/Echinobase/ PmDownload (pmin.scaf.fa)
Pelodiscus sinensis (Psi)	turtle	*PelSin_1.0*	Ensembl
Petromyzon marinus (Pma)	sea lamprey	*Pmarinus_7.0*	Ensembl
		germline genome	personal communication (Chris Amemiya, April 2016)
Pongo abelii (Pab)	orang utan	*PPYG2*	Ensembl
Procavia capensis (Pca)	hyrax	*proCap1*	Ensembl
Pteropus vampyrus (Pva)	megabat	*pteVam1*	Ensembl
Python molurus bivittatus (Pmo)	python	*Python 5.0.2*	NCBI/GCA_000186305.2
Rattus norvegicus (Rno)	rat	*Rnor_5.0*	Ensembl

Continued on next page

Table B.1 – continued from previous page

Latin name	Trivial name	Genome version	Genome source
Saccoglossus kowalevskii (Sko)	acorn worm	*JGI3.0*	ftp://ftp.jgi-psf.org/pub/compgen/metazome/v3.0/Skowalevskii/assembly/Saccoglossus_kowalevskii_v3.fasta
Sarcophilus harrisii (Sha)	Tasmanian devil	*Devil_ref v7.0*	`Ensembl`
Sorex araneus (Sar)	shrew	*sorAra2.0*	`Pre!Ensembl` (GCA_000181275.2)
Strongylocentrotus purpuratus (Spu)	purple sea urchin	*Spur_3.1*	`Ensembl Metazoa`
Struthio camelus (Sca)	ostrich	*v1.0*	NCBI/GCA_000698965.1
Sus scrofa (Ssc)	pig	*Sscrofa10.2*	`Ensembl`
Taeniopygia guttata (Tgu)	zebra finch	*taeGut3.2.4*	`Ensembl`
Takifugu rubripes (Tru)	pufferfish	*FUGU4*	`Ensembl`
Tupaia belangeri (Tbe)	tree shrew	*tupBel1*	`Ensembl`
Xenopus tropicalis (Xtr)	frog	*JGI 4.2*	`Ensembl`
Xiphophorus maculatus (Xma)	platyfish	*Xipmac4.4.2*	`Ensembl`

Table B.2: Cross-paralog conservation of arrestin isoforms according to public resources. Homologous isoforms that are known to exist in different orthology groups are marked in gray. Abbreviations: e – exon; * – known from literature; no mark – consideration of mouse (Mumu), human (Hsa), cow (Bta) `Ensembl` annotations, `http://www.proteinatlas.org`; ** – protein evidence for `Ensembl` isoform supported by Kim et al. (2014) or Ezkurdia et al. (2014) given in Uhlén et al. (2015)

Paralog	Isoform/Exon deviation	Resource	Species
*SAG***, *ARRB1***, *ARRB2***, *ARR3***	full-length	Ensembl, reference in most publications	Hsa, Mumu, Bta
*ARRB1***	alternate e1 (independent in both species)	Ensembl	Hsa, Mumu
*ARR3**	skipping of e3	Zhu et al. (2002)	Mumu
*ARRB2***, *ARRB1*	skipping of e4	Ensembl	*ARRB2* Hsa, Mumu; *ARRB1* Mumu
ARRB1	elongation of e4, frame shift	`Ensembl`	Mumu

Continued on next page

Table B.2 – continued from previous page

Paralog	Isoform/Exon deviation	Resource	Species
SAG	elongation and stop after e4	Ensembl	Hsa
ARR3	alternate start e5, frame shift, stop in e7	Ensembl	Hsa
*ARRB2***	elongated e5	Ensembl	Hsa
SAG	stop after e7	Ensembl	Hsa
*ARRB1***, *ARRB2***	start e8 (encodes only the *arrestin_C* domain)	Ensembl	Hsa
*ARRB1***	extension e10, frame shift, stop in e8	Ensembl	Hsa
ARRB2	start e9	Ensembl	Mumu
ARRB1	start e11	Ensembl	Mumu
*ARRB1***	stop after e11	Ensembl	Hsa
*ARRB2***	shortened e11, frame shift	Ensembl	Hsa
*SAG**	stop after e12	Palczewski and Smith (1996)	Mumu
*SAG**, *ARR3**	skipping of e12	Smith (1996) and Zhu et al. (2002)	*SAG* Hsa; *ARR3* Mumu
*ARRB1***, *ARR3**	skipping of e13	Parruti et al. (1993), Komori et al. (1998), and Zhu et al. (2002)	*ARRB1* Hsa, Bta, rat, cat; *ARR3* Mumu, Ensembl *ARRB1* Hsa, Mumu
*ARR3**	skipping of e13, 14	Zhu et al. (2002)	Mumu
ARR3	skipping of e15	Ensembl	Ensembl *ARR3* Hsa
ARRB2	elongated e14 (*ARRB2L*)	Sterne-Marr et al. (1993)	Bta, Ensembl Mumu
SAG	alternate, short e15	Ensembl	Ensembl *SAG* Bta
ARR3, *SAG*	alternate, short e16 (p44)	Smith et al. (1994)	Bta *SAG*, Ensembl Bta *SAG*, Ensembl *ARR3* Hsa
ARRB1	alternate, short e16	Ensembl	Mumu *ARRB1*

Table B.3: List of additional omics data considered for a refined annotation of arrestins. Latin and trivial names as well as acession numbers are given for data sets that were investigated on top of the `NCBI EST` and `TSA` database.

Latin name	Trivial name	GEO accession/ version	Transcriptome source
Callorhinchus milii	ghost shark	*GSM643959*	http://esharkgenome.imcb.a-star.edu.sg
Leucoraja erinacea	little skate	*GSM643957*	http://www.skatebase.org/downloads
Scyliorhinus canicula	catshark	*GSM643958*	http://www.skatebase.org/downloads
Gallus gallus	chicken	Carre et al. (2006)	http://www.chickest.udel.edu
Taeniopygia guttata	zebra finch	Jarvis et al. (2002)	http://songbird transcriptome.net/
		Replogle et al. (2008)	http://titan.biotec.uiuc.edu/cgi-bin/ESTWebsite/estima_start?seq Set=songbird3

Table B.5: Selection pressure acting on positively selected foreground branches of arrestin gene trees. Specific branches within the arrestin gene tree were tested for positive selection using the branch-site model of `codeml`, part of the `PAML` program (Fig. A.2). Inferred selection pressures (ω) and fraction of sites (p) in background and foreground branches are shown for the tests that rejected the null hypothesis (section 1.4.2, purifying and neutral selection at all positions in background and foreground branches). Abbreviation: Q – Quantile.

Foreground branch	Inferred parameter	μ	σ	Q_1	Q_3
ARR0.1 sea urchin	p_0	0.758	0.061	0.717	0.768
	p_1	0.094	0.031	0.073	0.097
	ω_0	0.036	0.007	0.031	0.037
	ω_1	40.036	144.328	2.629	11.827
ARR0.2 sea urchin	p_0	0.855	0.028	0.838	0.856
	p_1	0.096	0.034	0.074	0.098
	ω_0	0.034	0.009	0.028	0.034
	ω_1	202.310	326.497	31.409	56.590
SAG.1 ghost shark	p_0	0.696	0.042	0.663	0.693
	p_1	0.177	0.050	0.138	0.177
	ω_0	0.082	0.014	0.075	0.084
	ω_1	29.261	120.136	2.109	3.075
SAGa teleost	p_0	0.789	0.203	0.801	0.849
	p_1	0.038	0.023	0.023	0.036
	ω_0	0.075	0.010	0.067	0.074
	ω_1	775.970	408.641	999.000	999.000
SAGb teleost	p_0	0.831	0.274	0.892	0.928
	p_1	0.040	0.032	0.021	0.035
	ω_0	0.076	0.010	0.068	0.076
	ω_1	338.349	431.012	21.072	57.405
SAGb Acanthopterygii	p_0	0.838	0.127	0.817	0.863

Continued on next page

Table B.5 – continued from previous page

Foreground branch	Inferred parameter	μ	σ	Q_1	Q_3
	p_1	0.044	0.027	0.025	0.042
	ω_0	0.075	0.010	0.067	0.074
	ω_1	294.957	418.024	5.624	17.935
ARR3b euteleosts	p_0	0.716	0.096	0.679	0.729
	p_1	0.141	0.039	0.112	0.141
	ω_0	0.087	0.011	0.079	0.088
	ω_1	69.381	231.522	1.216	2.770

Table B.7: Specificity determining positions identified for different arrestin subgroups. This list shows all those residues that were found to be specificity determining and are not displayed in any of the sequence logos. The distance to known functional residues is given in parenthesis.

Paralog	Position in group	Position in cow	Functional annotation
ARR0.1	45	V43	close to receptor specificity residue (5)
ARR0.1	58	T56	neighboring to μ2 adaptin binding residue
ARR0.1	60	T58	neighboring to μ2 adaptin binding residue
ARR0.1	62	A60	close to receptor specificity residue (3), μ2 adaptin binding residue (3)
ARR0.1	85	N83	second neighboring to receptor specificity residue
ARR0.1	86	V84	second neighboring to receptor specificity residue
ARR0.1	105	E102	second neighboring to three element interaction
ARR0.1	106	R103	neighboring to three element interaction
ARR0.1	114	H111	close to three element interaction (3)
ARR0.1	206	H198	close to receptor specificity residue (5)
ARR0.1	215	I207	-
ARR0.1	217	Y209	-
ARR0.1	231	N222	close to receptor specificity residue (5), close to phosphodiesterase binding residues (3)
ARR0.1	396	P371	close to clathrin binding site (4)
ARR0.1	400	E378	clathrin binding site
ARR0.1	442	G409	-
SAG	30	H34	neighboring to polar core
SAG	31	V35	second neighboring to polar core
SAG	56	S60	neighboring to μ2 adaptin binding site
SAG	148	I149	-
SAG	152	A153	close to receptor specificity residue (4)
SAG	186	V190	close to receptor specificity residue (4)
SAG	218	V222	phosphodiesterase binding

Continued on next page

Table B.7 – continued from previous page

Paralog	Position in group	Position in cow	Functional annotation
SAG	225	S229	close to receptor specificity residue (4)
SAG	277	V281	-
SAG	311	G315	close to receptor specificity residue (3)
SAG	326	K330	IP6 binding (not conserved in *SAG*)
SAG	329	L333	neighboring to IP6 binding residue
ARRB2	42	V42	-
ARRB2	368	V369	close to clathrin binding site (3)
ARRB2	407	D406	-
ARRB2	408	Q407	-
ARR3	14	K16	phosphodiesterase binding
ARR3	24	F26	neighboring to polar core
ARR3	53	M55	neighboring to μ adaptin binding residue
ARR3	94	L97	close to three element interaction (3)
ARR3	98	Q101	neighboring to three element interaction
ARR3	100	R103	neighboring to three element interaction
ARR3	108	N111	close to three element interaction (3)
ARR3	116	M119	close to PxxP motif (5)
ARR3	147	C150	close to receptor specificity residue (4)
ARR3	179	G179	close to PxxP motif (3)
ARR3	218	V218	phosphodiesterase binding
ARR3	272	F272	close to receptor specificity residue (5)
ARR3	291	L289	neighboring to polar core, phosphodiesterase binding
ARR3	309	R307	close to receptor specificity residue (5)
ARR3	325	V323	neighboring to IP6 binding residue
ARR3	355	H353	-

Table B.8: Functional residues of arrestins considered in the current study. Furthermore the three element interaction, polar core and receptor binding residues were considered (listed in Tab. 1.3). The reference species is cow unless stated otherwise (mouse, *Mus musculus*, Mumu).

Study	Residues	Reference paralog	Implication
Vishnivetskiy et al. (2004)	K10, K11, R165, K170	Arr-2	phosphate sensor
Sutton et al. (2005)	R18	Arr-1	phosphate sensor
Hanson and Gurevich (2006)	K257	Arr-1	phosphate sensor
Benovic (1995)	R171, R175, K176, K166, K167	Arr-1	phosphate sensor
Milano et al. (2006)	K232, R236, K250, K324, K326	Arr-2	IP6 high affinity binding site

Continued on next page

Table B.8 – continued from previous page

Study	Residues	Reference paralog	Implication
Milano et al. (2006)	K157, K160, R161	Arr-2	IP6 low affinity binding site
Luttrell (1999)	P88xxP91, $p_1$21xx$p_1$24, $p_1$75xx$p_1$78	Arr-2	PxxP motif/c-Src binding
Schmid et al. (2006)	D385, F388, F391, R395	Arr-2	AP-2 β binding
Laporte et al. (2000)	R394, R396	Mumu Arr-3	AP-2 β binding
Marion et al. (2007)	Y54, L57	Arr-2	μ2 adaptin binding
Kang et al. (2009)	N367, L368, I369, E370, L371, D372	Arr-2 S	major clathrin binding site
Kang et al. (2009)	G333, L335, G336, D337, L338, S340	Arr-2	minor clathrin binding site
Baillie et al. (2007)	K18, T20, R26, R286, D291, L215-H220	Arr-3	phosphodiesterase binding
Szczepek et al. (2014)	67YGREDIDVMGL77	Arr-1	finger loop region (receptor binding)
Kang et al. (2015)	11-19, E71, D72, D74, M76, G77, L78, 79-86,Q134, D139, F197, F198, M199, L250, Y251, R319, T320, F339, L343	Mumu Arr-1	receptor binding
Zhan et al. (2011a)	I233, N245, M255, E256, A258, T261	Arr-2	receptor binding
Vishnivetskiy et al. (2011)	L48, G50, A51, Y238, C242, K250, C251, P252, M255, L68, S86, D240, D259, T261	Arr-2	receptor binding
Hanson.2006	78, V139, T157, L173, T233, S273, L77, F79, F85, F197	Arr-1	receptor binding
Zhan et al. (2016)	first 25AA	Arr-3	JNK3 activation
Seo et al. (2011)	V343, L278, S280, H350, D351, H352, I353	Arr-3	JNK3 activation
Kim et al. (2011)	F86, F196	Mumu Arr-1	oligomerization Arr-1

Table B.4: Model selection parameters during Bayesian inference of arrestin gene trees. The best model was identified with a path sampling approach implemented in the `BEAST2` model-selection app (Baele et al., 2012; Baele and Lemey, 2013). Parameters of the substitution models were determined by testing for the best models in `JModelTest` and `Prottest`, for nucleotide (NT) and amino acid (AA) alignments according to Akaike Information Content (AIC), respectively. The simple HKY was additionally tested and was the only NT model to converge during Markov Chain Monte Carlo (MCMC) sampling. The following priors were fixed: Birth-Death Model as tree prior with a uniform birth rate [0-10000], relaxed clock log normal as molecular clock. The best model for each alignment type, identified by pairwise testing with the Bayes Factor, is highlighted in gray.

Type	Partition	Substitution model parameters	Relative death rate parameters	Path sampling: chain length, steps	Marginal likelihood estimate
AA	no	JTT+I+G, γ: 1.04, p-inv: 0.05	$\alpha = 1, \beta = 10$	1 Mio., 100	-39388.156
AA	no	JTT+I+G, γ: 1.04, p-inv: 0.05	estimated α, β	1 Mio., 100	-39387.805
AA	no	JTT+I+G, estimated γ, substitution rates, p-inv	estimated α, β	1 Mio., 100	-39399.584
NT	no	GTR+I+G, γ: 0.85, p-inv: 0.09	$\alpha = 1, \beta = 10$	2 Mio., 50	-107773.801
NT	Codon	(1-3) HKY; estimated shape, substitution rates	$\alpha = 1, \beta = 10$	1 Mio., 100	-106680.711
NT	Codon	(1) GTR+I+G, p-inv: 0.09, γ: 1.14 (2) GTR+I+G, p-inv: 0.11, γ: 0.69 (3) TVM+I+G, p-inv: 0.0020, γ: 4.05	$\alpha = 1, \beta = 10$	3 Mio., 100	-106468.221
NT	Codon	as above	$\alpha = 1, \beta = 2$	2 Mio., 100	-106467.567

Table B.9: Fragments of *ARRB2* detected in birds. Table of potential *ARRB2* fragments detected by `tblastn` in 50 bird genomes. The respective genomes were queried with *SAG* from turkey, *ARRB1* from turtle, *ARRB2* from turtle and *ARR3* from finch (section 3.2.3). The number in the last column enumerates the number of exons retrieved additionally by querying the short read archive (`SRA`) with exons from very close relatives. The best *E*-value of the hit is shown, which was retrieved with any of the four queries. Species, in which more than two exons were found, are highlighted in gray. Abbreviation: e – exon.

Name	Contig	*E*-value	Exon	Length of contig	SRA
Acanthisitta chloris	*C15682494*	$1.00 \cdot 10^{-7}$	e12	119	
Anas platyrhynchos	-				

Continued on next page

Table B.9 – continued from previous page

Name	Contig	E-value	Exon	Length of contig	SRA
Antrostomus carolinensis	-/?				
Apaloderma vittatum	*C10759007*	1.5	part of e3	110	
Aptenodytes forster	*C12536524*	$8.00 \cdot 10^{-5}$	e16	141	
Apteryx mantelli	*NW_014005377.1*	$4.00 \cdot 10^{-7}$	e3, e4, e11	14,014	1
Aquila chrysaetos	-				7
Balearica regulorum	*C11911089*	$4.00 \cdot 10^{-5}$	part of e5	112	
	C11980919	$6.00 \cdot 10^{-8}$	e7	122	
Buceros rhinoceros	*C11575124*	$3.00 \cdot 10^{-4}$	e16	115	
Calypte anna	*scaffold171*	0.87	part of e11	1,249,265	
Cariama cristata	-				
Cathartes aura	-				
Chaetura pelagica	-				
Charadrius vociferus	*Scaffold3639*	0.002	e6	480	
Chlamydotis undulata	-/?				
Colius striatus	-/?				
Columba livia	*scaffold161*	6.1	e14	11,859,676	
Corvus brachyrhynchos	-/?				
Cuculus canorus	-/?				
Egretta garzetta	-/?				
Eurypyga helias	*C14487117*	$2.00 \cdot 10^{-7}$	part of e5	107	
Falco peregrinus	-				
Fulmarus glacialis	-/?				
Gallus gallus 4	-				
Gavia stellata	-				
Geospiza fortis	-				
Haliaeetus albicilla	-/?				
Haliaeetus leucocephalus	*Scaffold8956*	$2.00 \cdot 10^{-18}$	e7, e8, e16	2,789	
	Scaffold4072	$1.00 \cdot 10^{-17}$	e10, e11, e12, e14	70,001	
	Scaffold333272	0.058	part of e5	291	
Leptosomus discolor	*C10961796*	$8.00 \cdot 10^{-9}$	e13	100	
Manacus vitellinus	-				
Meleapris galloparo 5.0	-				
Melopsittacus undulatus	-				
Merops nubicus	-				
Mesitornis unicolor	-				
Nestor notabilis	-				
Nipponia nippon	*C13081413*	$6.00 \cdot 10^{-7}$	e7	170	
	C13300931	0.019	part of e9, e10	339	
	C13138967	0.3	e6	199	

Continued on next page

Table B.9 – continued from previous page

Name	**Contig**	***E*-value**	**Exon**	**Length of contig**	**SRA**
	C13365123		e3, e4	430	
Ophisthocomus hoazin	-				
Pelecanus crispus	-				
Phaethon lepturus	-				
Phalacrocorax carbo	-				
Phoenicopterus ruber	-				
Picoides pubescens	-				
Podiceps cristatus	-				
Pterocles gutturalis	-				
Pygoscelis adeliae	-				
Struthio camelus	*C14095491*	$6.00 \cdot 10^{-9}$	part of e11	154	1
	C14052965	$8.00 \cdot 10^{-6}$	e6	137	
	scaffold1684	$6.00 \cdot 10^{-5}$	e12	2,266	
Taeniopygia guttata	-				
Tauraco erythrolophus	-				
Tinamus major	*scaffold11991*	$4.00 \cdot 10^{-5}$	e10	2,426	
Tyto alba	-				

Table B.6: Analysis of natural selection after arrestin duplication. Specific branches within the arrestin gene tree were tested for positive selection using the branch-site model of `codeml`, part of the `PAML` program (Fig. A.2). The null hypothesis assumes purifying or neutral selection on the foreground and background branches, while the alternative model allows for positive selection on the foreground branch. Results of the likelihood ratio tests with associated P-values for the foreground branches that returned a significant P-value for any of the codon models. * < 0.05, ** < 0.01 ,*** < 0.001.

Codon model	**Foreground branch**	**Likelihood ratio**	P-**value**	**Significance level**
F3X4	*ARR0.1* sea urchin	3.770786	0.052	-
codon table	*ARR0.1* sea urchin	2.51134	0.113	-
F3X4	*ARR0.2* sea urchin	5.046812	0.025	*
codon table	*ARR0.2* sea urchin	5.448354	0.02	*
F3X4	*SAG.1* ghost shark	3.664146	0.056	-
codon table	*SAG.1* ghost shark	4.313492	0.038	*
F3X4	*SAGa* teleost	4.999824	0.025	*
codon table	*SAGa* teleost	2.142176	0.143	-
F3X4	*SAGb* teleost	7.37818	0.007	**
codon table	*SAGb* teleost	5.191498	0.023	*
F3X4	*SAGb* Acanthopterygii	4.782412	0.029	*
codon table	*SAGb* Acanthopterygii	3.18778	0.074	-
F3X4	*ARR3b* euteleost	4.72419	0.03	*
codon table	*ARR3b* euteleost	4.257344	0.039	*

Table B.10: Arrestin residues with post-translational modifications. The following post-translational modifications (PTMs) of arrestins were considered if investigated in literature or confirmed by phosphoproteomics: Hy – Hydroxyproline; Ni – Nitrosylation; Ph – Phosphorylation; Su – SUMOylation; Ub – Ubiquitination. Please see Tab. B.11 for the respective motifs and their conservation pattern and Tab. B.1 for species abbreviations. Other abbreviations: Arr – arrestin; CBS – clathrin binding site.

Study	Residue	Reference paralog	PTM	Function of PTM	Function of residue
Hoffert et al. (2006)	T231	Rno Arr-1	Ph	-	-
Hornbeck et al. (2015)	Y254	Hsa Arr-1	Ph	-	receptor specificity
Hornbeck et al. (2015)	Y258	Hsa Arr-1	Ph	-	receptor specificity
Cao et al. (2007), Weber, Schreiber, and Daub (2012), and Hornbeck et al. (2015)	Y47	Mumu, Rno, Hsa Arr-2	Ph	-	neighboring to receptor specificity position
Marion et al. (2007)	Y54	Hsa Arr-2	Ph	regulation of adaptin binding and endocytosis	μ2 adaptin binding site
Fernández-Arenas et al. (2014)	S163	Hsa Arr-2	Ph	receptor interaction	neighboring to phosphate sensor
Hornbeck et al. (2015)	Y173	Rno, Hsa Arr-2	Ph	-	neighboring to PxxP motif
Cassier et al. (2017)	T374	Rno Arr-2	Ph	-	neighboring to major CBS
Hoffert et al. (2006), Mertins et al. (2016), and Robles, Humphrey, and Mann (2017) a.o.	T410	Rno, Mumu, Hsa Arr-2	Ph	-	-
Lin et al. (1997), Lin et al. (1999), Barthet et al. (2009), and Huttlin et al. (2010)	S412	Mumu, Rno, Hsa Arr-2	Ph	prevents Src activation, dephosphorylation required for clathrin binding and endocytosis	-
Shenoy and Lefkowitz (2005) and Shenoy et al. (2009)	K11, K12	Hsa Arr-3	Ub	stability of receptor–arrestin interaction, activation of ERK, regulation of endosomal trafficking	phosphate sensor, receptor specificity
Paradis et al. (2015)	S14, T276	Hsa Arr-3	Ph	sequestration of CXCR4	receptor specificity
Shenoy and Lefkowitz (2005) and Mosser et al. (2008)	one of K18, 107, 108, 207, 296	Mumu Arr-3	Ub	regulation of receptor trafficking and degradation	phosphodiesterase binding, receptor specificity, phosphate sensor
Yan et al. (2011)	P176, P179, P181	Hsa Arr-3	Hy	inhibition of receptor internalization	PxxP motif
Ballif et al. (2008), Hornbeck et al. (2015), and Mertins et al. (2016)	Y48	Mumu, Rno, Hsa Arr-3	Ph	-	neighboring to receptor specificity position

Continued on next page

Table B.10 – continued from previous page

Study	Residue	Reference paralog	PTM	Function of PTM	Function of residue
Khoury et al. (2014)	T178	Rno Arr-3	Ph	increase of receptor interaction, trafficking and intracellular signaling	neighboring to PxxP motif
Kim et al. (2011) and Mertins et al. (2013)	K178	Hsa Arr-3	Ub	-	neighbor to PxxP motif
Cassier et al. (2017)	S194,267 or S268,281	Arr-3 Rno	Ph	-	receptor specificity; (second) neighboring to receptor specificity position; JNK activation
Choudhary et al. (2009) and Weintz et al. (2010)	S197	Mumu Arr-3	Ph	-	-
Wyatt et al. (2011)	K295, K400	Bta Arr-3	Su	necessary for endocytosis, possibly promotion of AP-2 binding	phosphate sensor; -
Xiao et al. (2015)	K295	Hsa Arr-3	Su	decreased binding and inhibition of ubiquitin ligase TRAF6 activation	phosphate sensor
Kim et al. (2002), Por et al. (2013), and Cassier et al. (2017)	T382/T383	Mumu, Btr, Hsa Arr-3	Ph	minor influence on trafficking, regulation of receptor interaction	-
Lin et al. (2002)	T383, S361	Rno Arr-3	Ph	regulation of clathrin-mediated internalization	-
Helou et al. (2013) and Hornbeck et al. (2015)	Y404	Hsa Arr-3	Ph	-	-
Ozawa et al. (2008)	C409	Hsa Arr-3	Ni	promotion of clathrin and/or adaptin binding	-

Table B.11: Pattern and conservation of arrestin post-translational modifications. Conservation of the respective enzyme motifs were evaluated based on the `human protein reference database` (HPRD) (Amanchy et al., 2007). Please see Tab. B.10 for the known function and the specific post-translational modification that is added at the respective position and Tab. B.1 for species abbreviations. Other abbreviations: CK – casein kinase; PKA/C – protein kinase A/C; GRK – G protein coupled receptor kinase; * – putative kinase based on HPRD scan; Arr – arrestin, Ati – salamander (*Ambystoma tigrinum*).

Study	Residue	Enzyme	Enzyme motif	Conservation pattern
Hoffert et al. (2006)	T231	PKA*/PKC*	[pS/pT]X[R/K]	fully conserved
Hornbeck et al. (2015)	Y254	Src*/ALK*	pY[A/G/S/T/E/D]; pYXXXX[F/Y]	conserved in all visual arrestins except Arr-4 Ati, Pmo, Ami, Arr-1 Aca, some teleosts; conserved in all visual arrestins except for Arr-4 Ati, Pmo, Ami
Hornbeck et al. (2015)	Y258	EGFR*/TC-PTP*	X[E/D]pYX; [E/D/Y]pY	lobe-finned fish Arr-1-specific Y/F except Lch; motif strictly conserved in mammals
Cao et al. (2007), Weber, Schreiber, and Daub (2012), and Hornbeck et al. (2015)	Y47	different kinases (ALK, EGFR)*	[E/D]XXpY, X[E/D]pYX	conserved in all Arr-0, -2, -3, -4 except Arr-4 teleosts
Marion et al. (2007)	Y54	Src, JAK2*	pYXX[L/I/V]	mammalian Arr-2-specific Y, otherwise F; Some phylogenetic groups of other paralogs also carry Y
Fernández-Arenas et al. (2014)	S163	PKC	[R/K]XX[pS/pT]	conserved in all arrestins except for primate Arr-4 (Y)
Hornbeck et al. (2015)	Y173	Src*	pY[A/G/S/T/E/D]	conserved in all Arr-2 except for lampreys; conserved in Actinoterygii Arr-3, Arr-0
Cassier et al. (2017)	T374	β-Adrenergic Receptor kinase*	[E/D][pS/pT]XXX	strictly conserved in Arr-2, mostly conserved in Arr-3 with few exceptions
Robles, Humphrey, and Mann (2017), Mertins et al. (2016), and Hoffert et al. (2006) a.o.	T410	CK1*	[E/D]XX[pS/pT]	tetrapod Arr-2-specific except marsupials
Barthet et al. (2009), Lin et al. (1997), Lin et al. (1999), and Huttlin et al. (2010)	S412	GRK5, ERK1/2	X[pS/pT]P	Arr-2-specific
Shenoy and Lefkowitz (2005) and Shenoy et al. (2009)	K11, K12	Mdm2, USP33	?	conserved in all Arr except for K11R in Arr-1 Ami, Ccu, Tgu, Pmo

Continued on next page

Table B.11 – continued from previous page

Study	Residue	Enzyme	Enzyme motif	Conservation pattern
Paradis et al. (2015)	S14, T276	ERK1/2	X[pS/pT]P	strictly conserved in Arr-2, -3, conserved in Arr-4 except for mammals, teleosts, conserved in most Arr0; Conserved in Arr-2, -3, -4
Shenoy and Lefkowitz (2005) and Mosser et al. (2008)	one of K18, K107, K108, K207, K296	?	?	-
Yan et al. (2011)	P176, P179, P181	prolyl hydroxylase 2	?	conserved in all Arr-2, -3 except for few exceptions, conserved in Arr-1 sauropsids, Arr-4 of most mammals
Ballif et al. (2008), Hornbeck et al. (2015), and Mertins et al. (2016)	Y48	different kinases (ALK, EGFR)*	[E/D]XXpY, X[E/D]pYX	see above
Khoury et al. (2014)	T178	ERK1/2	X[pS/pT]P	T/S in Arr-3 of single species: rodents, Sha, Meu, Amis, Tru, Ler; variable in Arr-1, -4
Kim et al. (2011) and Mertins et al. (2013)	K178	?	?	conserved in Arr-2, -3 except for Arr-2 lobe-finned fish and Lch, variable in Arr-1, -4, see above
Cassier et al. (2017)	S194, S267 or S268, S281	GRK1*; DNA-dependent protein kinase*; -*	X[pS/pT]XXX[A/P/S/T]; P[pS/pT]X	kinase pattern is rodent Arr-3-specific, residue is conserved in all arrestins; conserved in Arr-2, -3 except teleost Arr-3a, most Arr-1 except teleost Arr-1a; Mammalian Arr-3-specific, sauropsid Arr-4-specific
Choudhary et al. (2009) and Weintz et al. (2010)	S197	different kinases (CK1, PKA/C)*	[E/D]XX[pS/pT]; [R/K]X[pS/pT]	specific kinase-patterns are rodent Arr-3-specific, residue itself is Arr-3-specific
Wyatt et al. (2011)	K295, K400 (main site)	E2 ligase Ubc9	ΨsKxD/E	conserved in all arrestins with few exceptions; conserved in Arr-2, -3 except for Arr-2a + Arr-3a teleosts, Arr-3 Lca, Arr-2 Aca
Xiao et al. (2015)	K295	?	ΨsKxD/E	see above
Kim et al. (2002), Por et al. (2013), and Cassier et al. (2017)	T382/T383	CK2, MEK	[pS/pT]XX[E/D]	mammalian Arr-3-specific
Lin et al. (2002)	T383, S361	CK2, MEK	[pS/pT]XX[E/D]	see above; Arr-3-specific except Pmo, Lca
Helou et al. (2013) and Hornbeck et al. (2015)	Y404	different kinases (ALK, EGFR, Src)*	[E/D]XXpY; X[E/D]pYX; pY[A/G/S/T/E/D]	mammalian Arr-3-specific except Ssc, Aime, Rno, Mumu, Meu
Ozawa et al. (2008)	C409	?	?	Arr-3-specific except Arr-3b Tru

Table B.12: Cross-paralog conservation of arrestin isoforms accessed by conservation of reading frame and premature stop codons in deuterostomes. Conservation of isoforms was evaluated based on the updated annotation resulting from this study. The exceptions deviate from the pattern within the group. Please see Tab. B.1 for species abbreviations. Other abbreviations: e – exon; M – mammals; S – sauropsids; F – ray-finned fish.

Splice variant	***SAG***				***ARRB1***				***ARRB2***				***ARR3***				***ARR0***		
	M	S	F	Exceptions?	M	S	F	Exceptions?	M	S	F	Exceptions?	M	S	F	Exceptions?	Cin	Bfl	E
start within e8	x	x	x	Pmo, Aca, Gmo paralogs	x	x	x	-	x	x	x	Pma, Lca	-	x	x	Dre ARR3b	x	x	x except sea urchins
stop after e7	x	x	x	Gmo *SAGb*; Dre *SAGa*	x	x	x	Gac, Asme, Dre, Gga, Pmo, Xtr	x	x	x	Tru Asme, Gac *ARRB2b*; Dre, Nni *ARRB2a*; Pmo, Xtr, Lch	-	-	-	Gac, Ola, Oni, Xma *ARR3b*; Ami, Apl, Gga, Hle, Mga, Nni, Oho, Psi, Pmo, Sca, Tgu, Lch, Pma	x	x	e7, e8 fused in Sko, e7 split in other echinoderms
skipping of e4	x	x	x	-	x	x	x	Pma *ARRB1.2*	x	x	x	-	x	x	x	-	x	x	e4, 5 fused
skipping of e12	x	x	x	Gmo SAGb	x	x	x	-	x	x	x	-	x	x	x	Lca *ARRB2*; Otomorpha *ARR3b* e12, e13 fused	x	x	x
skipping of e13	x	x	x	-	x	x	x	exon missing in Nni, not possible in Pma *ARRB1.2*, no e13 in Pma *ARRB1.1*	-	-	-	-	x	x	x	Asme *ARR3b*, Dre *ARR3b*: e12, e13 fused	no e13	-	no e13
skipping of e15	x	x	-	exon not detected in Mdo, Lch, Xtro, Cmi *SAG.1*; Cmi *SAG.2* ends after e14; Psi, Aca ?	-	-	-	possible in Pma*ARRB1.1*	-	-	-	-	x	x	-	exon not detected in Ami, Cca, Lco, Psi, Nni, Pma; not possible in Lco	e14, e15 fused	x	x for Pmi, Sko

Appendix C

CV

Henrike Indrischek

Curriculum Vitae

January 2018

Address: Talstraße 12a, Zi 505
04103 Leipzig
Phone: 01728463570
Email: henrike@bioinf.uni-leipzig.de

Education

2013 - present	PhD candidate, Institute for Informatics, Universität Leipzig (submission planned for October 2017) title: High quality gene annotation for deep phylogenetic analysis supervisors: Prof. Peter F. Stadler, Prof. Sonja J. Prohaska
2013	M.Sc. Biochemistry, focus: Biomedicine, grade: 1.1 (A), Universität Leipzig
2011	B.Sc. Biochemistry, grade: 1.4 (A), Universität Leipzig
2007	Abitur, grade: 1.1 (A), Luther-Melanchthon Gymnasium, Lutherstadt Wittenberg

Research experience

2013 - present	Research associate, Bioinformatics Department, Universität Leipzig computational genomics, gene annotation, phylogenetics
05-08/2016	Research intern, Pharmacology Department, Vanderbilt University, US protein purification, western blot, experimental molecular biology
2013	Master student, Institute for Biochemistry and Bioinformatics Department, Universität Leipzig RNA purification and assays, analysis of next generation sequencing data, computational RNA folding
2012	Research assistent, Center for Biotechnology and Biomedicine, Universität Leipzig cell culture techniques, immuncytochemistry, fluorescence microscopy, atom force microscopy
2011	Research assistent, Paul Flechsig Institute for brain research, Universität Leipzig immunhistochemistry, enzyme activity assays
08-09/2010	Short-term research internship, Vanderbilt University, US blue native PAGE

Grants

2016 Micro grant awarded for a 3-month research internship at Vanderbilt University, US
2011 DAAD RISE stipend awarded for undergraduate internship at Vanderbilt University, US

Teaching experience

Supervision of undergraduate students during bachelor and master thesis projects.

2017	Instructor/ Teaching assistent	Phylogenomics module, Programming for Evolutionary biology workshop
2016	Teaching assistent	Programming for Evolutionary biology workshop
2015	Teaching assistent	Programming for Evolutionary biology workshop
2015	Teaching assistent	Sequence analysis and genomics, Bioinformatics master program
2014	Teaching assistent	Sequence analysis and genomics, Bioinformatics master program

Administrative experience

Member of organizing team "Bioinformatics fall seminar" (2014-2017).
This annual, five days Bioinformatics conference of the Bompfünewerer consortium with about 70 participants is organized by PhD students from the Bioinformatics Group and from the Computational EvoDevo Group, Universität Leipzig.

Relevant training

2016	Molecular Evolution workshop, Marine Biological Laboratory, Woods Hole, US
2015	Protein structure modeling with Rosetta, summer school by Prof. Jens Meiler, Universität Leipzig
2014	Training course LaTeX for students, Computer center, Universität Leipzig
2013	R for ecologists, Institute for Biology, Universität Leipzig
2008	English course, graduation with First Cambridge Certificate (grade: A), Language School International, Brisbane, Australia

Relevant skills

IT skills

Scripting languages	Perl, R
Working systems	Windows, Linux
Document preparation systems	LaTeX, Microsoft Office, Libre office

Languages

English	fluent
German	native speaker
French	good working knowledge

Publications

Refereed research papers

1. **Indrischek, H.**, S. J. Prohaska, V. V. Gurevich, E. V. Gurevich, and P. F. Stadler (2017). Uncovering missing pieces: duplication and deletion history of arrestins in deuterostomes. *BMC evolutionary biology* **17**(1), 163.
2. Hölzer, M., V. Krähling, F. Amman, E. Barth, S. H. Bernhart, V. A. O. Carmelo, M. Collatz, G. Doose, F. Eggenhofer, J. Ewald, J. Fallmann, L. M. Feldhahn, M. Fricke, J. Gebauer, A. J. Gruber, F. Hufsky, **H. Indrischek**, S. Kanton, J. Linde, N. Mostajo, R. Ochsenreiter, K. Riege, L. Rivarola-Duarte, A. H. Sahyoun, S. J. Saunders, S. E. Seemann, A. Tanzer, B. Vogel, S. Wehner, M. T. Wolfinger, R. Backofen, J. Gorodkin, I. Grosse, I. Hofacker, S. Hoffmann, C. Kaleta, P. F. Stadler, S. Becker, and M. Marz (2016). Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells. *Scientific reports* **6**, 34589.
3. **Indrischek, H.**, N. Wieseke, P. F. Stadler, and S. J. Prohaska (2016). The paralog-to-contig assignment problem: High quality gene models from fragmented assemblies. *Algorithms for Molecular Biology* **11**(1), 199.
4. Höfling, C., **H. Indrischek**, T. Höpcke, A. Waniek, H. Cynis, B. Koch, S. Schilling, M. Morawski, H.-U. Demuth, S. Roßner, and M. Hartlage-Rübsamen (2014). Mouse strain and brain

region-specific expression of the glutaminyl cyclases QC and isoQC. *International Journal of Developmental Neuroscience* **36**, 64–73.

Papers under revision

1. Lokits, A., **H. Indrischek**, J. Meiler, H. Hamm, and P. F. Stadler (2018). Tracing the evolution of the heterotrimeric G protein α subunit in Deuterostomia. *BMC evolutionary biology*.

In preparation

1. Kolora, S. R. R., A. Weigert, A. ZarifSaffari, S. Kehr, M. B. Walter Costa, **H. Indrischek**, G. Doose, M. Chintalapati, K. Lohse, J. Overmann, B. Bunk, C. Bleidorn, K. Henle, K. Nowick, P. F. Stadler, R. Faria, and M. Schlegel. Divergent evolution in the genomes of the closely-related lacertids, Lacerta viridis and L. bilineata and implications for speciation.

Attendance at conferences

15th Bioinformatics fall seminar 2017	Doubice, Czech Republic	
Central German GCB 2017	Leipzig, Germany	
32nd TBI winter seminar 2017	Bled, Slovenia	talk
14th Bioinformatics fall seminar 2016	Doubice, Czech Republic	talk
Programming for evolutionary biology conference 2016	Belgrade, Serbia	talk
13th Bioinformatics fall seminar 2015	Doubice, Czech Republic	talk
Central German GCB 2015	Halle (Saale), Germany	poster
ISMB and Student Council Symposium 2015	Dublin, Ireland	poster
GCB 2014	Bielefeld, Germany	poster
12th Bioinformatics fall seminar 2014	Doubice, Czech Republic	talk
ISMB and Student Council Symposium 2014	Boston, US	poster
29th TBI winter seminar 2014	Bled, Slovenia	talk
11th Bioinformatics fall seminar 2013	Doubice, Czech Republic	talk
28th TBI winter seminar 2013	Bled, Slovenia	

Leipzig, January 2018
Henrike Indrischek

Bibliography

Abascal, F., R. Zardoya, and D. Posada (2005). "ProtTest: selection of best-fit models of protein evolution". In: *Bioinformatics* 21.9, pp. 2104–2105. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti263.

Abdulaeva, G., P. A. Hargrave, and W. C. Smith (1995). "The sequence of arrestins from rod and cone photoreceptors in the frogs Rana catesbeiana and Rana pipiens. Localization of gene transcripts by reverse-transcription polymerase chain reaction on isolated photoreceptors". In: *European journal of biochemistry / FEBS* 234.2, pp. 437–442. ISSN: 0014-2956. DOI: 10.1111/j.1432-1033.1995.437_b.x.

Ahn, S. et al. (2004). "Reciprocal regulation of angiotensin receptor-activated extracellular signal-regulated kinases by beta-arrestins 1 and 2". In: *The Journal of biological chemistry* 279.9, pp. 7807–7811. ISSN: 0021-9258. DOI: 10.1074/jbc.C300443200.

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705.

Akerborg, O. et al. (2009). "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis". In: *Proceedings of the National Academy of Sciences* 106.14, pp. 5714–5719. ISSN: 1091-6490. DOI: 10.1073/pnas.0806251106.

Alberts, B. (2011). *Molekularbiologie der Zelle*. 5th ed. Weinheim: Wiley-VCH. ISBN: 3527323848.

Alexander, P. A. et al. (2009). "A minimal sequence code for switching protein structure and function". In: *Proceedings of the National Academy of Sciences* 106.50, pp. 21149–21154. ISSN: 1091-6490. DOI: 10.1073/pnas.0906408106.

Altenhoff, A. M. and C. Dessimoz (2012). "Inferring Orthology and Paralogy". In: *Evolutionary genomics*. Ed. by M. Anisimova. Vol. 855-856. Methods in molecular biology. New York: Humana Press, pp. 259–279. ISBN: 978-1-61779-582-4. DOI: 10.1007/978-1-61779-582-4_9.

Altenhoff, A. M. et al. (2012). "Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs". In: *PLoS computational biology* 8.5, e1002514. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002514.

Altenhoff, A. M. et al. (2016). "Standardized benchmarking in the quest for orthologs". In: *Nature methods* 13.5, pp. 425–430. ISSN: 1548-7105. DOI: 10.1038/nmeth.3830.

Altschul, S. F. et al. (1990). "Basic local alignment search tool". In: *Journal of molecular biology* 215.3, pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

Alvarez, C. E. (2008). "On the origins of arrestin and rhodopsin". In: *BMC evolutionary biology* 8, p. 222. ISSN: 1471-2148. DOI: 10.1186/1471-2148-8-222.

Amanchy, R. et al. (2007). "A curated compendium of phosphorylation motifs". In: *Nature biotechnology* 25.3, pp. 285–286. ISSN: 1087-0156. DOI: 10.1038/nbt0307-285.

Amann, B. et al. (2014). "True blue: S-opsin is widely expressed in different animal species". In: *Journal of animal physiology and animal nutrition* 98.1, pp. 32–42. ISSN: 1439-0396. DOI: 10.1111/jpn.12016.

Ambreen, S., F. Khalil, and A. A. Abbasi (2014). "Integrating large-scale phylogenetic datasets to dissect the ancient evolutionary history of vertebrate genome". In: *Molecular phylogenetics and evolution* 78, pp. 1–13. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2014.05.002.

Andreeva, A. et al. (2014). "SCOP2 prototype: a new approach to protein structure mining". In: *Nucleic acids research* 42, pp. D310–4. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1242.

Andreou, A. M. and N. Tavernarakis (2009). "SUMOylation and cell signalling". In: *Biotechnology journal* 4.12, pp. 1740–1752. ISSN: 1860-7314. DOI: 10.1002/biot.200900219.

Anisimova, M., J. P. Bielawski, and Z. Yang (2002). "Accuracy and power of bayes prediction of amino acid sites under positive selection". In: *Molecular biology and evolution* 19.6, pp. 950–958. ISSN: 0737-4038.

Anisimova, M., R. Nielsen, and Z. Yang (2003). "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites". In: *Genetics* 164.3, pp. 1229–1236. ISSN: 0016-6731.

Aubry, L., D. Guetta, and G. Klein (2009). "The arrestin fold: variations on a theme". In: *Current genomics* 10.2, pp. 133–142. ISSN: 1389-2029.

Aubry, L. and G. Klein (2013). "True arrestins and arrestin-fold proteins: a structure-based appraisal". In: *Progress in molecular biology and translational science* 118, pp. 21–56. ISSN: 1878-0814. DOI: 10.1016/B978-0-12-394440-5.00002-4.

Babenko, V. N. et al. (2004). "Prevalence of intron gain over intron loss in the evolution of paralogous gene families". In: *Nucleic acids research* 32.12, pp. 3724–3733. ISSN: 1362-4962. DOI: 10.1093/nar/gkh686.

Baele, G. and P. Lemey (2013). "Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency". In: *Bioinformatics* 29.16, pp. 1970–1979. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt340.

Baele, G. et al. (2012). "Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty". In: *Molecular biology and evolution* 29.9, pp. 2157–2167. ISSN: 0737-4038. DOI: 10.1093/molbev/mss084.

Baillie, G. S. et al. (2007). "Mapping binding sites for the PDE4D5 cAMP-specific phosphodiesterase to the N- and C-domains of beta-arrestin using spot-immobilized peptide arrays". In: *The Biochemical journal* 404.1, pp. 71–80. ISSN: 1470-8728. DOI: 10.1042/BJ20070005.

Baker, J. L. et al. (2016). "Functional Divergence of the Nuclear Receptor NR2C1 as a Modulator of Pluripotentiality During Hominid Evolution". In: *Genetics* 203.2, pp. 905–922. ISSN: 0016-6731. DOI: 10.1534/genetics.115.183889.

Ballif, B. A. et al. (2008). "Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain". In: *Journal of proteome research* 7.1, pp. 311–318. ISSN: 1535-3893. DOI: 10.1021/pr0701254.

Barash, Y. et al. (2010). "Deciphering the splicing code". In: *Nature* 465.7294, pp. 53–59. ISSN: 1476-4687. DOI: 10.1038/nature09000.

Barthet, G. et al. (2009). "β-arrestin1 phosphorylation by GRK5 regulates G protein-independent 5-HT4 receptor signalling". In: *The EMBO journal* 28.18, pp. 2706–2718. ISSN: 0261-4189. DOI: 10.1038/emboj.2009.215.

Bastian, F. et al. (2008). "Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species". In: *Data integration in the life sciences*. Ed. by A. Bairoch. Vol. 5109. Lecture notes in computer science. Lecture notes in bioinformatics. Berlin:

Springer, pp. 124–131. ISBN: 978-3-540-69828-9. DOI: 10.1007/978-3-540-69828-9_12.

Behe, M. J. (2010). "Experimental evolution, loss-of-function mutations, and the first rule of adaptive evolution". In: *The Quarterly review of biology* 85.4, pp. 419–445. ISSN: 0033-5770.

Benovic, J. L. (1995). "Visual Arrestin Binding to Rhodopsin". In: *The Journal of biological chemistry* 270.11, pp. 6010–6016. ISSN: 0021-9258. DOI: 10.1074/jbc.270.11.6010.

Bentrop, J. et al. (2001). "UV-light-dependent binding of a visual arrestin 1 isoform to photoreceptor membranes in a neuropteran (Ascalaphus) compound eye". In: *FEBS letters* 493.2-3, pp. 112–116. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(01)02287-6.

Betancur-R., R. et al. (2013). "The Tree of Life and a New Classification of Bony Fishes". In: *PLoS currents*. ISSN: 2157-3999. DOI: 10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

Bielawski, J. P., J. L. Baker, and J. Mingrone (2016). "Inference of Episodic Changes in Natural Selection Acting on Protein Coding Sequences via CODEML". In: *Current protocols in bioinformatics* 54, pp. 6.15.1–6.15.32. ISSN: 1934-3396. DOI: 10.1002/cpbi.2.

Bielawski, J. P. and Z. Yang (2003). "Maximum likelihood methods for detecting adaptive evolution after gene duplication". In: *Genome Evolution*. Ed. by A. Meyer and Y. van de Peer. Dordrecht: Springer, pp. 201–212. ISBN: 978-94-010-3957-4. DOI: 10.1007/978-94-010-0263-9_20.

– (2005). "Maximum Likelihood Methods for Detecting Adaptive Protein Evolution". In: *Statistical Methods in Molecular Evolution*. Ed. by R. Nielsen. Statistics for Biology and Health. New York: Springer, pp. 103–124. ISBN: 0-387-22333-9. DOI: 10.1007/0-387-27733-1_5.

Birney, E. (2000). "Using GeneWise in the Drosophila Annotation Experiment". In: *Genome research* 10.4, pp. 547–548. ISSN: 1088-9051. DOI: 10.1101/gr.10.4.547.

Blanchette, M. et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner". In: *Genome research* 14.4, pp. 708–715. ISSN: 1088-9051. DOI: 10.1101/gr.1933104.

Boardman, P. E. et al. (2002). "A Comprehensive Collection of Chicken cDNAs". In: *Current Biology* 12.22, pp. 1965–1969. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(02)01296-4.

Bockaert, J. and J. P. Pin (1999). "Molecular tinkering of G protein-coupled receptors: an evolutionary success". In: *The EMBO journal* 18.7, pp. 1723–1729. ISSN: 0261-4189. DOI: 10.1093/emboj/18.7.1723.

Bohn, L. M. et al. (1999). "Enhanced morphine analgesia in mice lacking beta-arrestin2". In: *Science (New York, N.Y.)* 286.5449, pp. 2495–2498. ISSN: 1095-9203.

Botero-Castro, F. et al. (2017). "Avian genomes revisited: Hidden genes uncovered and the rates vs. traits paradox in birds". In: *Molecular biology and evolution*. ISSN: 0737-4038. DOI: 10.1093/molbev/msx236.

Bouckaert, R. R. et al. (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 10.4, e1003537. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003537.

Braasch, I. and J. H. Postlethwait (2012). "Polyploidy in Fish and the Teleost Genome Duplication". In: *Polyploidy and genome evolution*. Ed. by P. S. Soltis and D. E. Soltis. Berlin and New York: Springer, pp. 341–384. ISBN: 9783642432811.

Brandt, B. W., K. A. Feenstra, and J. Heringa (2010). "Multi-Harmony: detecting functional specificity from sequence alignment". In: *Nucleic acids research* 38, W35–40. ISSN: 1362-4962. DOI: 10.1093/nar/gkq415.

– (2016). *multi-Harmony: multi-group Sequence Harmony & multi-Relief*. URL: http://www.ibi.vu.nl/programs/shmrwww/ (visited on 11/09/2016).

Brent, M. R. (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation". In: *Nature reviews. Genetics* 9.1, pp. 62–73. ISSN: 1471-0064. DOI: 10.1038/nrg2220.

Brown, B. M. et al. (2010). "Visual Arrestin 1 contributes to cone photoreceptor survival and light adaptation". In: *Investigative ophthalmology & visual science* 51.5, pp. 2372–2380. ISSN: 0146-0404. DOI: 10.1167/iovs.09-4895.

Brown, C. J., A. K. Johnson, and G. W. Daughdrill (2010). "Comparing models of evolution for ordered and disordered proteins". In: *Molecular biology and evolution* 27.3, pp. 609–621. ISSN: 0737-4038. DOI: 10.1093/molbev/msp277.

Brown, C. J. et al. (2002). "Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions". In: *Journal of molecular evolution* 55.1, pp. 104–110. ISSN: 1432-1432. DOI: 10.1007/s00239-001-2309-6.

Burge, C. B. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA". In: *Journal of molecular biology* 268.1, pp. 78–94. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.0951.

Burkard, R., M. Dell'Amico, and S. Martello (2012). *Assignment Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Bürkle, A. (2002). "Posttranslational Modification". In: *Encyclopedia of genetics*. Ed. by S. Brenner, J. H. Miller, and W. J. Broughton. San Diego: Academic Press, p. 1533. ISBN: 9780122270802. DOI: 10.1006/rwgn.2001.1022.

Burset, M., I. A. Seledtsov, and V. V. Solovyev (2001). "SpliceDB: database of canonical and non-canonical mammalian splice sites". In: *Nucleic acids research* 29.1, pp. 255–259. ISSN: 1362-4962.

Burtey, A. et al. (2007). "The conserved isoleucine-valine-phenylalteranine motif couples activation state and endocytic functions of beta-arrestins". In: *Traffic (Copenhagen, Denmark)* 8.7, pp. 914–931. ISSN: 1398-9219. DOI: 10.1111/j.1600-0854.2007.00578.x.

Byrum, C. A. et al. (2006). "Protein tyrosine and serine-threonine phosphatases in the sea urchin, Strongylocentrotus purpuratus: identification and potential functions". In: *Developmental biology* 300.1, pp. 194–218. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2006.08.050.

Cahill, T. J. et al. (2017). "Distinct conformations of GPCR-β-arrestin complexes mediate desensitization, signaling, and endocytosis". In: *Proceedings of the National Academy of Sciences* 114.10, pp. 2562–2567. ISSN: 1091-6490. DOI: 10.1073/pnas.1701529114.

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC bioinformatics* 10, p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421.

Cao, L. et al. (2007). "Quantitative time-resolved phosphoproteomic analysis of mast cell signaling". In: *Journal of immunology (Baltimore, Md. : 1950)* 179.9, pp. 5864–5876. ISSN: 0022-1767.

Carre, W. et al. (2006). "Chicken genomics resource: sequencing and annotation of 35,407 ESTs from single and multiple tissue cDNA libraries and CAP3 assembly of a chicken gene index". In: *Physiological genomics* 25.3, pp. 514–524. ISSN: 1094-8341. DOI: 10.1152/physiolgenomics.00207.2005.

Carugo, O. and F. Eisenhaber (2010). *Data mining techniques for the life sciences*. Vol. 609. Methods in molecular biology. New York, N.Y.: Humana Press. ISBN: 1603272402.

Cassier, E. et al. (2017). "Phosphorylation of β-arrestin2 at Thr(383) by MEK underlies β-arrestin-dependent activation of Erk1/2 by GPCRs". In: *eLife* 6. ISSN: 2050-084X. DOI: 10.7554/eLife.23777.

Celniker, G. et al. (2013). "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function". In: *Israel Journal of Chemistry* 53.3-4, pp. 199–206. ISSN: 0021-2148. DOI: 10.1002/ijch.201200096.

Chagoyen, M., J. A. García-Martín, and F. Pazos (2016). "Practical analysis of specificity-determining residues in protein families". In: *Briefings in bioinformatics* 17.2, pp. 255–261. ISSN: 1467-5463. DOI: 10.1093/bib/bbv045.

Chakraborty, A. and S. Chakrabarti (2015). "A survey on prediction of specificity-determining sites in proteins". In: *Briefings in bioinformatics* 16.1, pp. 71–88. ISSN: 1467-5463. DOI: 10.1093/bib/bbt092.

Chang, C.-T. et al. (2013). "Chtop is a component of the dynamic TREX mRNA export complex". In: *The EMBO journal* 32.3, pp. 473–486. ISSN: 0261-4189. DOI: 10.1038/emboj.2012.342.

Chen, S. et al. (1997). "Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes". In: *Neuron* 19.5, pp. 1017–1030. ISSN: 0896-6273.

Chen, W. et al. (2004). "Activity-dependent internalization of smoothened mediated by beta-arrestin 2 and GRK2". In: *Science (New York, N.Y.)* 306.5705, pp. 2257–2260. ISSN: 1095-9203. DOI: 10.1126/science.1104135.

Cheng, H. et al. (2006). "Human mRNA export machinery recruited to the 5' end of mRNA". In: *Cell* 127.7, pp. 1389–1400. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.10.044.

CHI (2017). *Discover on Target Brochure*. URL: http://www.discoveryontarget.com/GPCR-drug-discovery/.

Chi, P. B. and D. A. Liberles (2016). "Selection on protein structure, interaction, and sequence". In: *Protein science : a publication of the Protein Society* 25.7, pp. 1168–1178. ISSN: 1469-896X. DOI: 10.1002/pro.2886.

Chor, B. and T. Tuller (2005). "Maximum likelihood of evolutionary trees: hardness and approximation". In: *Bioinformatics* 21 Suppl 1, pp. i97–106. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti1027.

Choudhary, C. et al. (2009). "Mislocalized activation of oncogenic RTKs switches downstream signaling outcomes". In: *Molecular cell* 36.2, pp. 326–339. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2009.09.019.

Cleghorn, W. M. et al. (2015). "Arrestins regulate cell spreading and motility via focal adhesion dynamics". In: *Molecular biology of the cell* 26.4, pp. 622–635. ISSN: 1939-4586. DOI: 10.1091/mbc.E14-02-0740.

Cloonan, N. (2015). "Re-thinking miRNA-mRNA interactions: intertwining issues confound target discovery". In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 37.4, pp. 379–388. ISSN: 0265-9247. DOI: 10.1002/bies.201400191.

Coleman, J. E. and S. L. Semple-Rowland (2005). "GC1 deletion prevents light-dependent arrestin translocation in mouse cone photoreceptor cells". In: *Investigative ophthalmology & visual science* 46.1, pp. 12–16. ISSN: 0146-0404. DOI: 10.1167/iovs.04-0691.

Collins, B. M. et al. (2008). "Structure of Vps26B and mapping of its interaction with the retromer protein complex". In: *Traffic (Copenhagen, Denmark)* 9.3, pp. 366–379. ISSN: 1398-9219. DOI: 10.1111/j.1600-0854.2007.00688.x.

Corbo, J. C. et al. (2010). "CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors". In: *Genome research* 20.11, pp. 1512–1525. ISSN: 1088-9051. DOI: 10.1101/gr.109405.110.

Cortesi, F. et al. (2015). "Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1493–1498. ISSN: 1091-6490. DOI: 10.1073/pnas.1417803112.

Coulombe-Huntington, J. and J. Majewski (2007). "Characterization of intron loss events in mammals". In: *Genome research* 17.1, pp. 23–32. ISSN: 1088-9051. DOI: 10.1101/gr.5703406.

Craft, C. M. (2011). "Two are better than one: unraveling the functions of cone arrestin in zebrafish (Commentary on Renninger, Gesemann and Neuhauss)". In: *The European journal of neuroscience* 33.4, p. 657. ISSN: 0953-816X. DOI: 10.1111/j.1460-9568.2011.07625.x.

Craft, C. M. and D. H. Whitmore (1995). "The arrestin superfamily: cone arrestins are a fourth family". In: *FEBS letters* 362.2, pp. 247–255. ISSN: 0014-5793. DOI: 10.1016/0014-5793(95)00213-S.

Craft, C. M., D. H. Whitmore, and A. F. Wiechmann (1994). "Cone arrestin identified by targeting expression of a functional family". In: *The Journal of biological chemistry* 269.6, pp. 4613–4619. ISSN: 0021-9258.

Craft, C. M. et al. (2014). "Primate Short-Wave cones share molecular markers with rods". In: *Retinal Degenerative Diseases: Mechanisms and Experimental Therapy*. Ed. by J. Ash et al. New York: Springer, pp. 49–56. ISBN: 9781461432098. DOI: 10.1007/978-1-4614-3209-8_7.

Crisp, A. et al. (2015). "Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes". In: *Genome biology* 16, p. 50. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0607-3.

Crooks, G. E. et al. (2004). "WebLogo: a sequence logo generator". In: *Genome research* 14.6, pp. 1188–1190. ISSN: 1088-9051. DOI: 10.1101/gr.849004.

Cunningham, F. et al. (2015). "Ensembl 2015". In: 43, pp. D662–9. DOI: 10.1093/nar/gku1010.

Curwen, V. et al. (2004). "The Ensembl Automatic Gene Annotation System". In: *Genome research* 14.5, pp. 942–950. ISSN: 1088-9051. DOI: 10.1101/gr.1858004.

Dalquen, D. A. et al. (2012). "ALF–a simulation framework for genome evolution". In: *Molecular biology and evolution* 29.4, pp. 1115–1123. ISSN: 0737-4038. DOI: 10.1093/molbev/msr268.

Dang, C. C. et al. (2010). "FLU, an amino acid substitution model for influenza proteins". In: *BMC evolutionary biology* 10.1, p. 99. ISSN: 1471-2148. DOI: 10.1186/1471-2148-10-99.

Darriba, D. et al. (2011). "ProtTest 3: fast selection of best-fit models of protein evolution". In: *Bioinformatics* 27.8, pp. 1164–1165. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr088.

– (2012). "jModelTest 2: more models, new heuristics and parallel computing". In: *Nature methods* 9.8, p. 772. ISSN: 1548-7105. DOI: 10.1038/nmeth.2109.

Davies, W. I. L., M. W. Hankins, and R. G. Foster (2010). "Vertebrate ancient opsin and melanopsin: divergent irradiance detectors". In: *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology* 9.11, pp. 1444–1457. ISSN: 1474-9092. DOI: 10.1039/c0pp00203h.

Davies, W. I. L. et al. (2009). "Into the blue: gene duplication and loss underlie color vision adaptations in a deep-sea chimaera, the elephant shark Callorhinchus milii".

In: *Genome research* 19.3, pp. 415–426. ISSN: 1088-9051. DOI: 10.1101/gr.084509.108.

Dayhoff, M. O., R. M. Schwartz, and Orcutt BC (1978). "A Model of Evolutionary Change in Proteins". In: *Atlas of Protein Sequence and Structure* 3.5, pp. 345–352.

Dean, A. M. et al. (2002). "The Pattern of Amino Acid Replacements in alpha/beta-Barrels". In: *Molecular biology and evolution* 19.11, pp. 1846–1864. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a004009.

Delroisse, J. et al. (2014). "High opsin diversity in a non-visual infaunal brittle star". In: *BMC genomics* 15, p. 1035. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-1035.

Deming, J. D. et al. (2015a). "Arrestin 1 and Cone Arrestin 4 Have Unique Roles in Visual Function in an All-Cone Mouse Retina". In: *Investigative ophthalmology & visual science* 56.13, pp. 7618–7628. ISSN: 0146-0404. DOI: 10.1167/iovs.15-17832.

Deming, J. D. et al. (2015b). "Visual Cone Arrestin 4 Contributes to Visual Function and Cone Health". In: *Investigative ophthalmology & visual science* 56.9, p. 5407. ISSN: 0146-0404. DOI: 10.1167/iovs.15-16647.

Derrien, T., R. Guigó, and R. Johnson (2011). "The Long Non-Coding RNAs: A New (P)layer in the Dark Matter". In: *Frontiers in genetics* 2, p. 107. ISSN: 1664-8021. DOI: 10.3389/fgene.2011.00107.

Dessimoz, C. et al. (2011). "Comparative genomics approach to detecting split-coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes)". In: *Briefings in bioinformatics* 12.5, pp. 474–484. ISSN: 1467-5463. DOI: 10.1093/bib/bbr038.

Dorval, K. M. et al. (2006). "CHX10 targets a subset of photoreceptor genes". In: *The Journal of biological chemistry* 281.2, pp. 744–751. ISSN: 0021-9258. DOI: 10.1074/jbc.M509470200.

Dos Santos, G. et al. (2014). "FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations". In: *Nucleic acids research*. ISSN: 1362-4962. DOI: 10.1093/nar/gku1099.

Drummond, A. J. and R. R. Bouckaert (2015). *Bayesian Evolutionary Analysis with BEAST*. Cambridge: Cambridge University Press. ISBN: 1139095110.

Du, C. and X. Xie (2012). "G protein-coupled receptors as therapeutic targets for multiple sclerosis". In: *Cell research* 22.7, pp. 1108–1128. ISSN: 1748-7838. DOI: 10.1038/cr.2012.87.

Echave, J., S. J. Spielman, and C. O. Wilke (2016). "Causes of evolutionary rate variation among protein sites". In: *Nature reviews. Genetics* 17.2, p. 109. ISSN: 1471-0064. DOI: 10.1038/nrg.2015.18.

Eddy, S. R. (1998). "Profile hidden Markov models". In: *Bioinformatics* 14.9, pp. 755–763. ISSN: 1367-4803.

– (2008). "A probabilistic model of local sequence alignment that simplifies statistical significance estimation". In: *PLoS computational biology* 4.5, e1000069. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000069.

– (2009). "A new generation of homology search tools based on probabilistic inference". In: *Genome informatics. International Conference on Genome Informatics* 23.1, pp. 205–211. ISSN: 0919-9454.

– (2011). "Accelerated Profile HMM Searches". In: *PLoS computational biology* 7.10, e1002195. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002195.

Eddy, S. R., T. J. Wheeler, and HMMER development team (2015). *HMMER User's Guide: Biological sequence analysis using profile hidden Markov models*. URL: http://eddylab.org/software/hmmer3/3.1b2/Userguide.pdf (visited on 11/01/2017).

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic acids research* 32.5, pp. 1792–1797. ISSN: 1362-4962. DOI: 10.1093/nar/gkh340.

Efron, B., E. Halloran, and S. Holmes (1996). "Bootstrap confidence levels for phylogenetic trees". In: *Proceedings of the National Academy of Sciences* 93.23, p. 13429. ISSN: 1091-6490. URL: http://www.pnas.org/content/93/23/13429.full.

Eid, J. et al. (2009). "Real-time DNA sequencing from single polymerase molecules". In: *Science (New York, N.Y.)* 323.5910, pp. 133–138. ISSN: 1095-9203. DOI: 10.1126/science.1162986.

Eilbeck, K. et al. (2009). "Quantitative measures for the management and comparison of annotated genomes". In: *BMC bioinformatics* 10, p. 67. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-67.

Eirin-Lopez, J. M. et al. (2004). "Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of orphon H1 genes". In: *Molecular biology and evolution* 21.10, pp. 1992–2003. ISSN: 0737-4038. DOI: 10.1093/molbev/msh213.

Ejsmond, M. J. and J. Radwan (2015). "Red Queen Processes Drive Positive Selection on Major Histocompatibility Complex (MHC) Genes". In: *PLoS computational biology* 11.11, e1004627. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004627.

Elkon, R., A. P. Ugalde, and R. Agami (2013). "Alternative cleavage and polyadenylation: extent, regulation and function". In: *Nature reviews. Genetics* 14.7, pp. 496–506. ISSN: 1471-0064. DOI: 10.1038/nrg3482.

Esnault, C., J. Maestre, and T. Heidmann (2000). "Human LINE retrotransposons generate processed pseudogenes". In: *Nature genetics* 24.4, pp. 363–367. ISSN: 1061-4036. DOI: 10.1038/74184.

Ezkurdia, I. et al. (2014). "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes". In: *Human Molecular Genetics* 23.22, pp. 5866–5878. ISSN: 1460-2083. DOI: 10.1093/hmg/ddu309.

Fablet, M. et al. (2009). "Evolutionary origin and functions of retrogene introns". In: *Molecular biology and evolution* 26.9, pp. 2147–2156. ISSN: 0737-4038. DOI: 10.1093/molbev/msp125.

Fahim, A. T., S. P. Daiger, and R. G. Weleber (1993). "Nonsyndromic Retinitis Pigmentosa Overview". In: *GeneReviews((R))*. Ed. by M. P. Adam et al. Seattle (WA).

Falcon, J. et al. (2014). "Drastic neofunctionalization associated with evolution of the timezyme AANAT 500 Mya". In: *Proceedings of the National Academy of Sciences* 111.1, pp. 314–319. ISSN: 1091-6490. DOI: 10.1073/pnas.1312634110.

Feenstra, K. A. et al. (2007). "Sequence harmony: detecting functional specificity from alignments". In: *Nucleic acids research* 35, W495–8. ISSN: 1362-4962. DOI: 10.1093/nar/gkm406.

Felsenstein, J. (1983). "Statistical Inference of Phylogenies". In: *Journal of the Royal Statistical Society. Series A (General)* 146.3, p. 246. ISSN: 0035-9238. DOI: 10.2307/2981654.

– (2017). *PHYLIP*. URL: http://evolution.genetics.washington.edu/phylip/ (visited on 09/20/2017).

Fernández-Arenas, E. et al. (2014). "β-Arrestin-1 mediates the TCR-triggered rerouting of distal receptors to the immunological synapse by a PKC-mediated mechanism". In: *The EMBO journal* 33.6, pp. 559–577. ISSN: 0261-4189. DOI: 10.1002/embj.201386022.

Feuda, R. et al. (2012). "Metazoan opsin evolution reveals a simple route to animal vision". In: *Proceedings of the National Academy of Sciences* 109.46, pp. 18868–18872. ISSN: 1091-6490. DOI: 10.1073/pnas.1204609109.

Finn, R. D. et al. (2014). "Pfam: the protein families database". In: *Nucleic acids research* 42, pp. D222–30. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1223.

Finn, R. D. et al. (2015). "HMMER web server: 2015 update". In: *Nucleic acids research* 43.W1, W30–8. ISSN: 1362-4962. DOI: 10.1093/nar/gkv397.

Fitch, W. M. (2000). "Homology: A personal view on some of the problems". In: *Trends in genetics : TIG* 16.5, pp. 227–231. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(00)02005-9.

Fitzpatrick, D. A., D. M. O'Halloran, and A. M. Burnell (2006). "Multiple lineage specific expansions within the guanylyl cyclase gene family". In: *BMC evolutionary biology* 6, p. 26. ISSN: 1471-2148. DOI: 10.1186/1471-2148-6-26.

Fletcher, W. and Z. Yang (2010). "The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection". In: *Molecular biology and evolution* 27.10, pp. 2257–2267. ISSN: 0737-4038. DOI: 10.1093/molbev/msq115.

Flicek, P. et al. (2014). "Ensembl 2014". In: *Nucleic acids research* 42, pp. D749–55. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1196.

Force, A. et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations". In: *Genetics* 151.4, pp. 1531–1545. ISSN: 0016-6731.

Foster, M. W., D. T. Hess, and J. S. Stamler (2009). "Protein S-nitrosylation in health and disease: a current perspective". In: *Trends in molecular medicine* 15.9, pp. 391–404. ISSN: 1471-499X. DOI: 10.1016/j.molmed.2009.06.007.

Franzosa, E. A. and Y. Xia (2009). "Structural determinants of protein evolution are context-sensitive at the residue level". In: *Molecular biology and evolution* 26.10, pp. 2387–2395. ISSN: 0737-4038. DOI: 10.1093/molbev/msp146.

Fried, C., S. J. Prohaska, and P. F. Stadler (2003). "Independent Hox-cluster duplications in lampreys". In: *Journal of experimental zoology. Part B, Molecular and developmental evolution* 299.1, pp. 18–25. ISSN: 1552-5007. DOI: 10.1002/jez.b.37.

Fritzell, K. et al. (2017). "ADARs and editing: The role of A-to-I RNA modification in cancer progression". In: *Seminars in cell & developmental biology*. ISSN: 1084-9521. DOI: 10.1016/j.semcdb.2017.11.018.

Furukawa, T., J. B. Hurley, and S. Kawamura, eds. (2014). *Vertebrate photoreceptors: Functional molecular bases*. Tokyo u.a.: Springer. ISBN: 978-4-431-54879-9.

Gaidarov, I. et al. (1999). "Arrestin function in G protein-coupled receptor endocytosis requires phosphoinositide binding". In: *The EMBO journal* 18.4, pp. 871–881. ISSN: 0261-4189. DOI: 10.1093/emboj/18.4.871.

Garcia-Fernandez, J. (2005). "The genesis and evolution of homeobox gene clusters". In: *Nature reviews. Genetics* 6.12, pp. 881–892. ISSN: 1471-0064. DOI: 10.1038/nrg1723.

Gascuel, O. and M. A. Steel (2007). *Reconstructing evolution: New mathematical and computational advances*. ISBN: 9780199208227. URL: http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=e000xat&AN=205558.

Gentles, A. J. et al. (2007). "Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica". In: *Genome research* 17.7, pp. 992–1004. ISSN: 1088-9051. DOI: 10.1101/gr.6070707.

Gerstein, M. B. et al. (2007). "What is a gene, post-ENCODE? History and updated definition". In: *Genome research* 17.6, pp. 669–681. ISSN: 1088-9051. DOI: 10.1101/gr.6339607.

Gharib, W. H. and M. Robinson-Rechavi (2013). "The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC". In: *Molecular biology and evolution* 30.7, pp. 1675–1686. ISSN: 0737-4038. DOI: 10.1093/molbev/mst062.

Gilad, Y. et al. (2003). "Natural Selection on the Olfactory Receptor Gene Family in Humans and Chimpanzees". In: *The American Journal of Human Genetics* 73.3, pp. 489–501. ISSN: 0002-9297. DOI: 10.1086/378132.

Gimenez, L. E. et al. (2012). "Manipulation of very few receptor discriminator residues greatly enhances receptor specificity of non-visual arrestins". In: *The Journal of biological chemistry* 287.35, pp. 29495–29505. ISSN: 0021-9258. DOI: 10.1074/jbc.M112.366674.

Girnita, L. et al. (2007). "Beta-arrestin and Mdm2 mediate IGF-1 receptor-stimulated ERK activation and cell cycle progression". In: *The Journal of biological chemistry* 282.15, pp. 11329–11338. ISSN: 0021-9258. DOI: 10.1074/jbc.M611526200.

Glasauer, S. M. K. and S. C. F. Neuhauss (2014). "Whole-genome duplication in teleost fishes and its evolutionary consequences". In: *Molecular genetics and genomics : MGG* 289.6, pp. 1045–1060. ISSN: 1617-4623. DOI: 10.1007/s00438-014-0889-2.

Goldman, N., J. L. Thorne, and D. T. Jones (1998). "Assessing the impact of secondary structure and solvent accessibility on protein evolution". In: *Genetics* 149.1, pp. 445–458. ISSN: 0016-6731. URL: http://europepmc.org/articles/PMC1460119?pdf=render.

Goldman, N. and Z. Yang (1994). "A codon-based model of nucleotide substitution for protein-coding DNA sequences". In: *Molecular biology and evolution* 11.5, pp. 725–736. ISSN: 0737-4038.

Gomez, M. D. P. et al. (2011). "Arrestin in ciliary invertebrate photoreceptors: molecular identification and functional analysis in vivo". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31.5, pp. 1811–1819. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3320-10.2011.

Goodman, O. B. et al. (1996). "Beta-arrestin acts as a clathrin adaptor in endocytosis of the beta2-adrenergic receptor". In: *Nature* 383.6599, pp. 447–450. ISSN: 1476-4687. DOI: 10.1038/383447a0.

Goodsell, D. S. (2003). "Src Tyrosine Kinase". In: *RCSB Protein Data Bank*. DOI: 10.2210/rcsb_pdb/mom_2003_7.

Gotoh, O. (2008). "Direct mapping and alignment of protein sequences onto genomic sequence". In: *Bioinformatics* 24.21, pp. 2438–2444. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn460.

Granzin, J. et al. (1998). "X-ray crystal structure of arrestin from bovine rod outer segments". In: *Nature* 391.6670, pp. 918–921. ISSN: 1476-4687. DOI: 10.1038/36147.

Granzin, J. et al. (2015). "Structural evidence for the role of polar core residue Arg175 in arrestin activation". In: *Scientific Reports* 5, p. 15808. ISSN: 2045-2322. DOI: 10.1038/srep15808.

Gremme, G. et al. (2005). "Engineering a software tool for gene structure prediction in higher organisms". In: *Information and Software Technology* 47.15, pp. 965–978. ISSN: 0950-5849. DOI: 10.1016/j.infsof.2005.09.005.

Guindon, S., O. Gascuel, and B. Rannala (2003). "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood". In: *Systematic Biology* 52.5, pp. 696–704. ISSN: 1076-836X. DOI: 10.1080/10635150390235520.

Guindon, S. et al. (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". In: *Systematic Biology* 59.3, pp. 307–321. ISSN: 1076-836X. DOI: 10.1093/sysbio/syq010.

Gulko, B. et al. (2015). "A method for calculating probabilities of fitness consequences for point mutations across the human genome". In: *Nature genetics* 47.3, ng.3196. ISSN: 1061-4036. DOI: 10.1038/ng.3196.

Gurevich, E. V., J. L. Benovic, and V. V. Gurevich (2002). "Arrestin2 and arrestin3 are differentially expressed in the rat brain during postnatal development". In: *Neuroscience* 109.3, pp. 421–436. ISSN: 0306-4522. DOI: 10.1016/s0306-4522(01)00511-5.

– (2004). "Arrestin2 expression selectively increases during neural differentiation". In: *Journal of neurochemistry* 91.6, pp. 1404–1416. ISSN: 0022-3042. DOI: 10.1111/j.1471-4159.2004.02830.x.

Gurevich, E. V. and V. V. Gurevich (2006a). "Arrestins: ubiquitous regulators of cellular signaling pathways". In: *Genome biology* 7.9, p. 236. ISSN: 1465-6906. DOI: 10.1186/gb-2006-7-9-236.

Gurevich, V. V. and E. V. Gurevich (2006b). "The structural basis of arrestin-mediated regulation of G-protein-coupled receptors". In: *Pharmacology & Therapeutics* 110.3, pp. 465–502. ISSN: 0163-7258. DOI: 10.1016/j.pharmthera.2005.09.008.

Gurevich, V. V. et al. (1995). "Arrestin Interactions with G Protein-coupled Receptors". In: *The Journal of biological chemistry* 270.2, pp. 720–731. ISSN: 0021-9258. DOI: 10.1074/jbc.270.2.720.

Gurevich, V. V. et al. (2011). "The functional cycle of visual arrestins in photoreceptor cells". In: *Progress in retinal and eye research* 30.6, pp. 405–430. ISSN: 1873-1635. DOI: 10.1016/j.preteyeres.2011.07.002.

Gurevich, V. V. et al. (2014). "Enhanced phosphorylation-independent arrestins and gene therapy". In: *Handbook of experimental pharmacology* 219, pp. 133–152. ISSN: 0171-2004. DOI: 10.1007/978-3-642-41199-1_7.

Hammesfahr, B. et al. (2015). *Scipio eukaryotic gene identification: Help*. URL: http://www.webscipio.org/help/webscipio#setting.

Han, M. et al. (2001). "Crystal Structure of β-Arrestin at 1.9 Å". In: *Structure* 9.9, pp. 869–880. ISSN: 0969-2126. DOI: 10.1016/S0969-2126(01)00644-X.

Hankins, M. W., W. I. L. Davies, and R. G. Foster (2014). "The Evolution of Non-visual Photopigments in the Central Nervous System of Vertebrates". In: *Evolution of Visual and Non-visual Pigments*. Ed. by D. M. Hunt et al. Vol. 4. Springer Series in Vision Research. Boston, MA and s.l.: Springer, pp. 65–103. ISBN: 978-1-4614-4355-1. DOI: 10.1007/978-1-4614-4355-1_3.

Hanson, S. M. and V. V. Gurevich (2006). "The differential engagement of arrestin surface charges by the various functional forms of the receptor". In: *The Journal of biological chemistry* 281.6, pp. 3458–3462. ISSN: 0021-9258. DOI: 10.1074/jbc.M512148200.

Hanson, S. M. et al. (2006). "Differential interaction of spin-labeled arrestin with inactive and active phosphorhodopsin". In: *Proceedings of the National Academy of Sciences* 103.13, pp. 4900–4905. ISSN: 1091-6490. DOI: 10.1073/pnas.0600733103.

Hanson, S. M. et al. (2008). "A model for the solution structure of the rod arrestin tetramer". In: *Structure* 16.6, pp. 924–934. ISSN: 0969-2126. DOI: 10.1016/j.str.2008.03.006.

Hao, H. et al. (2012). "Transcriptional regulation of rod photoreceptor homeostasis revealed by in vivo NRL targetome analysis". In: *PLoS Genetics* 8.4, e1002649. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002649.

Harty, B. L. et al. (2015). "Defining the gene repertoire and spatiotemporal expression profiles of adhesion G protein-coupled receptors in zebrafish". In: *BMC genomics* 16, p. 62. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1296-8.

Hasegawa, M., H. Kishino, and T.-a. Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of molecular evolution* 22.2, pp. 160–174. ISSN: 1432-1432. DOI: 10.1007/BF02101694.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1, p. 97. ISSN: 0006-3444. DOI: 10.2307/2334940.

Hatje, K. and M. Kollmar (2011). "Predicting Tandemly Arrayed Gene Duplicates with WebScipio". In: *Gene Duplication*. Ed. by F. Friedberg. InTech. ISBN: 978-953-307-387-3. DOI: 10.5772/24240.

Hatje, K. et al. (2011). "Cross-species protein sequence and gene structure prediction with fine-tuned Webscipio 2.0 and Scipio". In: *BMC research notes* 4, p. 265. ISSN: 1756-0500. DOI: 10.1186/1756-0500-4-265.

Haug-Baltzell, A. et al. (2015). "Identification of dopamine receptors across the extant avian family tree and analysis with other clades uncovers a polyploid expansion among vertebrates". In: *Frontiers in neuroscience* 9, p. 361. ISSN: 1662-4548. DOI: 10.3389/fnins.2015.00361.

He, X. and J. Zhang (2005). "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution". In: *Genetics* 169.2, pp. 1157–1164. ISSN: 0016-6731. DOI: 10.1534/genetics.104.037051.

Heled, J. and A. J. Drummond (2008). "Bayesian inference of population size history from multiple loci". In: *BMC evolutionary biology* 8.1, p. 289. ISSN: 1471-2148. DOI: 10.1186/1471-2148-8-289.

Hellmuth, M. et al. (2015). "Phylogenomics with paralogs". In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2058–2063. ISSN: 1091-6490. DOI: 10.1073/pnas.1412770112.

Helou, Y. A. et al. (2013). "ERK positive feedback regulates a widespread network of tyrosine phosphorylation sites across canonical T cell signaling and actin cytoskeletal proteins in Jurkat T cells". In: *PloS one* 8.7, e69641. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0069641.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks". In: *Proceedings of the National Academy of Sciences* 89.22, pp. 10915–10919. ISSN: 1091-6490. URL: http://www.pnas.org/content/89/22/10915.full.pdf.

Hirsch, J. A. et al. (1999). "A Model for Arrestin's Regulation: The 2.8 Å Crystal Structure of Visual Arrestin". In: *Cell* 97.2, pp. 257–269. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)80735-7.

Hisatomi, O. et al. (1997). "Arrestins expressed in killifish photoreceptor cells". In: *FEBS letters* 411.1, pp. 12–18. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(97)00640-6.

Hoeppner, C. Z., N. Cheng, and R. D. Ye (2012). "Identification of a nuclear localization sequence in β-arrestin-1 and its functional implications". In: *The Journal of biological chemistry* 287.12, pp. 8932–8943. ISSN: 0021-9258. DOI: 10.1074/jbc.M111.294058.

Hoffert, J. D. et al. (2006). "Quantitative phosphoproteomics of vasopressin-sensitive renal cells: regulation of aquaporin-2 phosphorylation at two sites". In: *Proceedings of the National Academy of Sciences* 103.18, pp. 7159–7164. ISSN: 1091-6490. DOI: 10.1073/pnas.0600895103.

Hoffmann, F. G., J. C. Opazo, and J. F. Storz (2010). "Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates". In: *Proceedings of the National Academy of Sciences* 107.32, pp. 14274–14279. ISSN: 1091-6490. DOI: 10.1073/pnas.1006756107.

Höhna, S., M. Defoin-Platel, and A. J. Drummond (2008). "Clock-constrained tree proposal operators in Bayesian phylogenetic inference". In: *8th IEEE International*

Conference on BioInformatics and BioEngineering. Ed. by K. S. Nikita. Piscataway, NJ: IEEE, pp. 1–7. ISBN: 978-1-4244-2844-1. DOI: 10.1109/BIBE.2008.4696663.

Holland, P. W. H. et al. (2017). "New genes from old: asymmetric divergence of gene duplicates and the evolution of development". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1713. ISSN: 0962-8436. DOI: 10.1098/rstb.2015.0480.

Hordijk, W. and O. Gascuel (2005). "Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood". In: *Bioinformatics* 21.24, pp. 4338–4347. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti713.

Horie, T., H. Orii, and M. Nakagawa (2005). "Structure of ocellus photoreceptors in the ascidian Ciona intestinalis larva as revealed by an anti-arrestin antibody". In: *Journal of neurobiology* 65.3, pp. 241–250. ISSN: 0022-3034. DOI: 10.1002/neu.20197.

Horie, T. et al. (2008). "Pigmented and nonpigmented ocelli in the brain vesicle of the ascidian larva". In: *The Journal of comparative neurology* 509.1, pp. 88–102. ISSN: 1096-9861. DOI: 10.1002/cne.21733.

Horita, H. et al. (2012). "Specialized motor-driven dusp1 expression in the song systems of multiple lineages of vocal learning birds". In: *PloS one* 7.8, e42173. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0042173.

Hornbeck, P. V. et al. (2015). "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations". In: *Nucleic acids research* 43, pp. D512–20. ISSN: 1362-4962. DOI: 10.1093/nar/gku1267. (Visited on 08/28/2017).

Hoskins, R. A. et al. (2007). "Sequence finishing and mapping of Drosophila melanogaster heterochromatin". In: *Science (New York, N.Y.)* 316.5831, pp. 1625–1628. ISSN: 1095-9203. DOI: 10.1126/science.1139816.

Huang, S.-P., B. M. Brown, and C. M. Craft (2010). "Visual Arrestin 1 acts as a modulator for N-ethylmaleimide-sensitive factor in the photoreceptor synapse". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30.28, pp. 9381–9391. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.1207-10.2010.

Huerta-Cepas, J., F. Serra, and P. Bork (2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Molecular biology and evolution* 33.6, pp. 1635–1638. ISSN: 0737-4038. DOI: 10.1093/molbev/msw046.

Hughes, A. L. (1994). "The evolution of functionally novel proteins after gene duplication". In: *Proceedings. Biological sciences / The Royal Society* 256.1346, pp. 119–124. ISSN: 1471-2954. DOI: 10.1098/rspb.1994.0058.

Huminiecki, L. and C. H. Heldin (2010). "2R and remodeling of vertebrate signal transduction engine". In: *BMC biology* 8, p. 146. ISSN: 1741-7007. DOI: 10.1186/1741-7007-8-146.

Huson, D. H. and C. Scornavacca (2012). "Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks". In: *Systematic Biology* 61.6, pp. 1061–1067. ISSN: 1076-836X. DOI: 10.1093/sysbio/sys062.

Huson, D. H. et al. (2007). "Dendroscope: An interactive viewer for large phylogenetic trees". In: *BMC bioinformatics* 8.1, pp. 1–6. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-460.

Huttlin, E. L. et al. (2010). "A tissue-specific atlas of mouse protein phosphorylation and expression". In: *Cell* 143.7, pp. 1174–1189. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.12.001.

Illergård, K., D. H. Ardell, and A. Elofsson (2009). "Structure is three to ten times more conserved than sequence–a study of structural response in protein cores". In: *Proteins* 77.3, pp. 499–508. ISSN: 0887-3585. DOI: 10.1002/prot.22458.

Illumina Inc. (2017). *Illumina sequencing platforms*. URL: https://www.illumina.com/systems/sequencing-platforms.html (visited on 12/08/2017).

Imanishi, Y., O. Hisatomi, and F. Tokunaga (1999). "Two types of arrestins expressed in medaka rod photoreceptors". In: *FEBS letters* 462.1-2, pp. 31–36. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(99)01483-0.

Indrischek, H. et al. (2016). "The paralog-to-contig assignment problem: High quality gene models from fragmented assemblies". In: *Algorithms for Molecular Biology* 11.1, p. 199. ISSN: 1748-7188. DOI: 10.1186/s13015-016-0063-y.

Indrischek, H. et al. (2017). "Uncovering missing pieces: duplication and deletion history of arrestins in deuterostomes". In: *BMC evolutionary biology* 17.1, p. 163. ISSN: 1471-2148. DOI: 10.1186/s12862-017-1001-4.

Ingram, N. T., A. P. Sampath, and G. L. Fain (2016). "Why are rods more sensitive than cones?" In: *The Journal of physiology* 594.19, pp. 5415–5426. ISSN: 1469-7793. DOI: 10.1113/JP272556.

Innan, H. and F. A. Kondrashov (2010). "The evolution of gene duplications: classifying and distinguishing between models". In: *Nature reviews. Genetics* 11.2, pp. 97–108. ISSN: 1471-0064. DOI: 10.1038/nrg2689.

Inoue, J. et al. (2015). "Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling". In: *Proceedings of the National Academy of Sciences* 112.48, pp. 14918–14923. ISSN: 1091-6490. DOI: 10.1073/pnas.1507669112.

Iwata, H. and O. Gotoh (2012). "Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features". In: *Nucleic acids research* 40.20, e161. ISSN: 1362-4962. DOI: 10.1093/nar/gks708.

Jarvis, E. D. et al. (2002). "A framework for integrating the songbird brain". In: *Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology* 188.11-12, pp. 961–980. ISSN: 0340-7594. DOI: 10.1007/s00359-002-0358-y.

Jean-Charles, P.-Y., V. Rajiv, and S. K. Shenoy (2016). "Ubiquitin-Related Roles of β-Arrestins in Endocytic Trafficking and Signal Transduction". In: *Journal of cellular physiology* 231.10, pp. 2071–2080. ISSN: 1097-4652. DOI: 10.1002/jcp.25317.

Jeffreys, H. (1935). "Some Tests of Significance, Treated by the Theory of Probability". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 31.02, p. 203. ISSN: 0305-0041. DOI: 10.1017/S030500410001330X.

Johnson, G. L. and R. Lapadat (2002). "Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases". In: *Science (New York, N.Y.)* 298.5600, pp. 1911–1912. ISSN: 1095-9203. DOI: 10.1126/science.1072682.

Johnsson, P., K. V. Morris, and D. Grandér (2014). "Pseudogenes: a novel source of trans-acting antisense RNAs". In: *Methods in molecular biology (Clifton, N.J.)* 1167, pp. 213–226. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-0835-6_14.

Jones, D. T., W. R. Taylor, and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences". In: *Bioinformatics* 8.3, pp. 275–282. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/8.3.275.

Jonkers, I. and J. T. Lis (2015). "Getting up to speed with transcription elongation by RNA polymerase II". In: *Nature reviews. Molecular cell biology* 16.3, pp. 167–177. ISSN: 1471-0072. DOI: 10.1038/nrm3953.

Juan, D. de, F. Pazos, and A. Valencia (2013). "Emerging methods in protein co-evolution". In: *Nature reviews. Genetics* 14.4, pp. 249–261. ISSN: 1471-0064. DOI: 10.1038/nrg3414.

Jukes, T. H. and C. R. Cantor (2013). "Evolution of Protein Molecules". In: *Mammalian Protein Metabolism*. Ed. by H. N. Munro. Burlington: Elsevier Science, pp. 21–132. ISBN: 9781483232119. DOI: 10.1016/B978-1-4832-3211-9.50009-7.

Jun, J. et al. (2009). "Duplication mechanism and disruptions in flanking regions determine the fate of Mammalian gene duplicates". In: *Journal of computational biology : a journal of computational molecular cell biology* 16.9, pp. 1253–1266. ISSN: 1066-5277. DOI: 10.1089/cmb.2009.0074.

Kaessmann, H., N. Vinckenbosch, and M. Long (2009). "RNA-based gene duplication: mechanistic and evolutionary insights". In: *Nature reviews. Genetics* 10.1, pp. 19–31. ISSN: 1471-0064. DOI: 10.1038/nrg2487.

Kang, D. S. et al. (2009). "Structure of an arrestin2-clathrin complex reveals a novel clathrin binding domain that modulates receptor trafficking". In: *The Journal of biological chemistry* 284.43, pp. 29860–29872. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.023366.

Kang, Y. et al. (2015). "Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser". In: *Nature* 523.7562, pp. 561–567. ISSN: 1476-4687. DOI: 10.1038/nature14656.

Kapustin, Y. et al. (2008). "Splign: algorithms for computing spliced alignments with identification of paralogs". In: *Biology direct* 3, p. 20. ISSN: 1745-6150. DOI: 10.1186/1745-6150-3-20.

Karlin, S. and L. Brocchieri (1996). "Evolutionary conservation of RecA genes in relation to protein structure and function". In: *Journal of bacteriology* 178.7, pp. 1881–1894. ISSN: 0021-9193.

Karp, R. M. (1972). "Reducibility among combinatorial problems". In: *Complexity of Computer Computations*. Ed. by R. E. Miller and J. W. Thatcher. Berkley, CA: Plenum, pp. 85–103.

Kass, R. E. and A. E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795. ISSN: 0162-1459. DOI: 10.1080/01621459.1995.10476572.

Kassahn, K. S. et al. (2009). "Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates". In: *Genome research* 19.8, pp. 1404–1418. ISSN: 1088-9051. DOI: 10.1101/gr.086827.108.

Kawano-Yamashita, E., M. Koyanagi, and A. Terakita (2014). "The Evolution and Diversity of Pineal and Parapineal Photopigments". In: *Evolution of Visual and Non-visual Pigments*. Ed. by D. M. Hunt et al. Vol. 4. Springer Series in Vision Research. Boston, MA and s.l.: Springer, pp. 1–21. ISBN: 978-1-4614-4355-1. DOI: 10.1007/978-1-4614-4355-1_1.

Kawano-Yamashita, E. et al. (2011). "beta-arrestin functionally regulates the non-bleaching pigment parapinopsin in lamprey pineal". In: *PloS one* 6.1, e16402. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0016402.

Keilwagen, J. et al. (2016). "Using intron position conservation for homology-based gene prediction". In: *Nucleic acids research* 44.9, e89. ISSN: 1362-4962. DOI: 10.1093/nar/gkw092.

Keller, O. et al. (2008). "Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species". In: *BMC bioinformatics* 9, p. 278. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-278.

Keller, O. et al. (2011). "A novel hybrid gene prediction method employing protein multiple sequence alignments". In: *Bioinformatics* 27.6, pp. 757–763. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr010.

Kern, R. C., D. S. Kang, and J. L. Benovic (2009). "Arrestin2/clathrin interaction is regulated by key N- and C-terminal regions in arrestin2". In: *Biochemistry* 48.30, pp. 7190–7200. ISSN: 0006-2960. DOI: 10.1021/bi900369c.

Keshava Prasad, T. S. et al. (2009). "Human Protein Reference Database–2009 update". In: *Nucleic acids research* 37, pp. D767–72. ISSN: 1362-4962. DOI: 10.1093/nar/gkn892.

Khoury, E. et al. (2014). "Differential regulation of endosomal GPCR/β-arrestin complexes and trafficking by MAPK". In: *The Journal of biological chemistry* 289.34, pp. 23302–23317. ISSN: 0021-9258. DOI: 10.1074/jbc.M114.568147.

Kim, J. et al. (2008). "Beta-arrestins regulate atherosclerosis and neointimal hyperplasia by controlling smooth muscle cell proliferation and migration". In: *Circulation research* 103.1, pp. 70–79. ISSN: 1524-4571. DOI: 10.1161/CIRCRESAHA.108.172338.

Kim, M.-S. et al. (2014). "A draft map of the human proteome". In: *Nature* 509.7502, pp. 575–581. ISSN: 1476-4687. DOI: 10.1038/nature13302.

Kim, M. et al. (2011). "Robust self-association is a common feature of mammalian visual arrestin-1". In: *Biochemistry* 50.12, pp. 2235–2242. ISSN: 0006-2960. DOI: 10.1021/bi1018607.

Kim, Y. J. et al. (2013). "Crystal structure of pre-activated arrestin p44". In: *Nature* 497.7447, pp. 142–146. ISSN: 1476-4687. DOI: 10.1038/nature12133.

Kim, Y.-M. et al. (2002). "Regulation of arrestin-3 phosphorylation by casein kinase II". In: *The Journal of biological chemistry* 277.19, pp. 16837–16846. ISSN: 0021-9258. DOI: 10.1074/jbc.M201379200.

Kimchi-Sarfaty, C. et al. (2007). "A silent polymorphism in the MDR1 gene changes substrate specificity". In: *Science (New York, N.Y.)* 315.5811, pp. 525–528. ISSN: 1095-9203. DOI: 10.1126/science.1135308.

Kimple, M. E. et al. (2014). "Inhibitory G proteins and their receptors: emerging therapeutic targets for obesity and diabetes". In: *Experimental & molecular medicine* 46, e102. ISSN: 2092-6413. DOI: 10.1038/emm.2014.40.

Kimura, A. et al. (2000). "Both PCE-1/RX and OTX/CRX Interactions Are Necessary for Photoreceptor-specific Gene Expression". In: *The Journal of biological chemistry* 275.2, pp. 1152–1160. ISSN: 0021-9258. DOI: 10.1074/jbc.275.2.1152.

Kirchhausen, T., D. Owen, and S. C. Harrison (2014). "Molecular structure, function, and dynamics of clathrin-mediated membrane traffic". In: *Cold Spring Harbor perspectives in biology* 6.5, a016725. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a016725.

Kiser, P. D. et al. (2012). "Key enzymes of the retinoid (visual) cycle in vertebrate retina". In: *Biochimica et biophysica acta* 1821.1, pp. 137–151. ISSN: 0006-3002. DOI: 10.1016/j.bbalip.2011.03.005.

Klein, D. C. (2006). "Evolution of the vertebrate pineal gland: the AANAT hypothesis". In: *Chronobiology international* 23.1-2, pp. 5–20. ISSN: 0742-0528. DOI: 10.1080/07420520500545839.

Klotz, E. and A. M. Newman (2013). "Practical guidelines for solving difficult linear programs". In: *Surveys in Operations Research and Management Science* 18.1-2, pp. 1–17. ISSN: 1876-7354. DOI: 10.1016/j.sorms.2012.11.001.

Koepfli, K.-P., B. Paten, and S. J. O'Brien (2015). "The Genome 10K Project: a way forward". In: *Annual review of animal biosciences* 3, pp. 57–111. ISSN: 2165-8102. DOI: 10.1146/annurev-animal-090414-014900.

Kohout, T. A. et al. (2001). "beta-Arrestin 1 and 2 differentially regulate heptahelical receptor signaling and trafficking". In: *Proceedings of the National Academy of Sciences* 98.4, pp. 1601–1606. ISSN: 1091-6490. DOI: 10.1073/pnas.041608198.

Kojima, K. et al. (2017). "Evolutionary steps involving counterion displacement in a tunicate opsin". In: *Proceedings of the National Academy of Sciences* 114.23, pp. 6028–6033. ISSN: 1091-6490. DOI: 10.1073/pnas.1701088114.

Komori, N. et al. (1998). "Differential expression of alternative splice variants of beta-arrestin-1 and -2 in rat central nervous system and peripheral tissues". In: *The European journal of neuroscience* 10.8, pp. 2607–2616. ISSN: 0953-816X.

Komori, N. et al. (1994). "Phosrestin I, an arrestin homolog that undergoes light-induced phosphorylation in dipteran photoreceptors". In: *Insect biochemistry and molecular biology* 24.6, pp. 607–617. ISSN: 0965-1748. DOI: 10.1016/0965-1748(94)90097-3.

Kondrashov, F. A. et al. (2002). "Selection in the evolution of gene duplications". In: *Genome biology* 3.2, research0008.1. ISSN: 1465-6906. DOI: 10.1186/gb-2002-3-2-research0008.

Kook, S., V. V. Gurevich, and E. V. Gurevich (2014). "Arrestins in apoptosis". In: *Handbook of experimental pharmacology* 219, pp. 309–339. ISSN: 0171-2004. DOI: 10.1007/978-3-642-41199-1_16.

Korenbrot, J. I. (2012). "Speed, sensitivity, and stability of the light response in rod and cone photoreceptors: facts and models". In: *Progress in retinal and eye research* 31.5, pp. 442–466. ISSN: 1873-1635. DOI: 10.1016/j.preteyeres.2012.05.002.

Korlach, J. et al. (2017). "De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads". In: *GigaScience* 6.10, pp. 1–16. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix085.

Kosiol, C. and N. Goldman (2005). "Different versions of the Dayhoff rate matrix". In: *Molecular biology and evolution* 22.2, pp. 193–199. ISSN: 0737-4038. DOI: 10.1093/molbev/msi005.

Koudritsky, M. and E. Domany (2008). "Positional distribution of human transcription factor binding sites". In: *Nucleic acids research* 36.21, pp. 6795–6805. ISSN: 1362-4962. DOI: 10.1093/nar/gkn752.

Koyanagi, M. et al. (2017). "Vertebrate Bistable Pigment Parapinopsin: Implications for Emergence of Visual Signaling and Neofunctionalization of Non-visual Pigment". In: *Frontiers in Ecology and Evolution* 5, p. 8083. ISSN: 2296-701X. DOI: 10.3389/fevo.2017.00023.

Krishnan, A. et al. (2015). "Evolutionary hierarchy of vertebrate-like heterotrimeric G protein families". In: *Molecular phylogenetics and evolution* 91, pp. 27–40. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2015.05.009.

Kristaponyte, I. et al. (2012). "Role of rhodopsin and arrestin phosphorylation in retinal degeneration of Drosophila". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32.31, pp. 10758–10766. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0565-12.2012.

Kroeber, S., C. Schomerus, and H.-W. Korf (1998). "A specific and sensitive double-immunofluorescence method for the demonstration of S-antigen and serotonin in trout and rat pinealocytes by means of primary antibodies from the same donor species". In: *Histochemistry and cell biology* 109.4, pp. 309–317. ISSN: 0948-6143. DOI: 10.1007/s004180050231.

Krupnick, J. G. and J. L. Benovic (1998). "The role of receptor kinases and arrestins in G protein-coupled receptor regulation". In: *Annual review of pharmacology and toxicology* 38, pp. 289–319. ISSN: 0362-1642. DOI: 10.1146/annurev.pharmtox.38.1.289.

Kuhrt, H. et al. (2017). "The Retina of Asian and African Elephants: Comparison of Newborn and Adult". In: *Brain, behavior and evolution* 89.2, pp. 84–103. ISSN: 1421-9743. DOI: 10.1159/000464097.

Kumar, A. (2015). "Bayesian phylogeny analysis of vertebrate serpins illustrates evolutionary conservation of the intron and indels based six groups classification system from lampreys for ∼500 MY". In: *PeerJ* 3, e1026. ISSN: 2167-8359. DOI: 10.7717/peerj.1026.

Kumar, A. et al. (2011). "Spliceosomal intron insertions in genome compacted ray-finned fishes as evident from phylogeny of MC receptors, also supported by a few other GPCRs". In: *PloS one* 6.8, e22046. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022046.

Kumar, S. et al. (2017). "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times". In: *Molecular biology and evolution* 34.7, pp. 1812–1819. ISSN: 0737-4038. DOI: 10.1093/molbev/msx116.

Kuo, F.-T., T.-L. Lu, and H.-W. Fu (2006). "Opposing effects of beta-arrestin1 and beta-arrestin2 on activation and degradation of Src induced by protease-activated receptor 1". In: *Cellular signalling* 18.11, pp. 1914–1923. ISSN: 1873-3913. DOI: 10.1016/j.cellsig.2006.02.009.

Kuraku, S., A. Meyer, and S. Kuratani (2009). "Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after?" In: *Molecular biology and evolution* 26.1, pp. 47–59. ISSN: 0737-4038. DOI: 10.1093/molbev/msn222.

La Garcia de Serrana, D., E. A. Mareco, and I. A. Johnston (2014). "Systematic variation in the pattern of gene paralog retention between the teleost superorders Ostariophysi and Acanthopterygii". In: *Genome biology and evolution* 6.4, pp. 981–987. ISSN: 1759-6653. DOI: 10.1093/gbe/evu074.

Lagman, D. et al. (2012). "Expansion of transducin subunit gene families in early vertebrate tetraploidizations". In: *Genomics* 100.4, pp. 203–211. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2012.07.005.

Lagman, D. et al. (2013). "The vertebrate ancestral repertoire of visual opsins, transducin alpha subunits and oxytocin/vasopressin receptors was established by duplication of their shared genomic region in the two rounds of early vertebrate genome duplications". In: *BMC evolutionary biology* 13, p. 238. ISSN: 1471-2148. DOI: 10.1186/1471-2148-13-238.

Lagman, D. et al. (2015). "Transducin duplicates in the zebrafish retina and pineal complex: differential specialisation after the teleost tetraploidisation". In: *PloS one* 10.3, e0121330. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0121330.

Lamb, T. D. et al. (2016). "Evolution of Vertebrate Phototransduction: Cascade Activation". In: *Molecular biology and evolution* 33.8, pp. 2064–2087. ISSN: 0737-4038. DOI: 10.1093/molbev/msw095.

Lander, E. S. et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062.

Laporte, S. A. et al. (2000). "The interaction of beta-arrestin with the AP-2 adaptor is required for the clustering of beta 2-adrenergic receptor into clathrin-coated pits". In: *The Journal of biological chemistry* 275.30, pp. 23120–23126. ISSN: 0021-9258. DOI: 10.1074/jbc.M002581200.

Laporte, S. A. et al. (2002). "beta-Arrestin/AP-2 interaction in G protein-coupled receptor internalization: identification of a beta-arrestin binging site in beta 2-adaptin". In: *The Journal of biological chemistry* 277.11, pp. 9247–9254. ISSN: 0021-9258. DOI: 10.1074/jbc.M108490200.

Lappano, R. and M. Maggiolini (2017). "Pharmacotherapeutic Targeting of G Protein-Coupled Receptors in Oncology: Examples of Approved Therapies and Emerging Concepts". In: *Drugs* 77.9, pp. 951–965. ISSN: 0012-6667. DOI: 10.1007/s40265-017-0738-9.

Laranjeiro, R. and D. Whitmore (2014). "Transcription factors involved in retinogenesis are co-opted by the circadian clock following photoreceptor differentiation". In: *Development (Cambridge, England)* 141.13, pp. 2644–2656. ISSN: 1477-9129. DOI: 10.1242/dev.104380.

Larhammar, D., K. Nordström, and T. A. Larsson (2009). "Evolution of vertebrate rod and cone phototransduction genes". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364.1531, pp. 2867–2880. ISSN: 0962-8436. DOI: 10.1098/rstb.2009.0077.

Lartillot, N. and H. Philippe (2006). "Computing Bayes factors using thermodynamic integration". In: *Systematic Biology* 55.2, pp. 195–207. ISSN: 1076-836X. DOI: 10.1080/10635150500433722.

Le, S. Q. and O. Gascuel (2008). "An improved general amino acid replacement matrix". In: *Molecular biology and evolution* 25.7, pp. 1307–1320. ISSN: 0737-4038. DOI: 10.1093/molbev/msn067.

Lechner, M. et al. (2011). "Proteinortho: detection of (co-)orthologs in large-scale analysis". In: *BMC bioinformatics* 12, p. 124. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-124.

Lees, J. G. et al. (2016). "Functional innovation from changes in protein domains and their combinations". In: *Current opinion in structural biology* 38, pp. 44–52. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2016.05.016.

Lesk, A. M. (2014). *Introduction to bioinformatics*. 4. ed. Oxford a.o.: Oxford University Press. ISBN: 978-0-19-965156-6.

Levy, E. D., S. De, and S. A. Teichmann (2012). "Cellular crowding imposes global constraints on the chemistry and evolution of proteomes". In: *Proceedings of the National Academy of Sciences* 109.50, pp. 20461–20466. ISSN: 1091-6490. DOI: 10.1073/pnas.1209312109.

Li, L., C. J. Stoeckert, and D. S. Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes". In: *Genome research* 13.9, pp. 2178–2189. ISSN: 1088-9051. DOI: 10.1101/gr.1224503.

Li, Y. I. and R. R. Copley (2013). "Scaffolding low quality genomes using orthologous protein sequences". In: *Bioinformatics* 29.2, pp. 160–165. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts661.

Lin, F.-T. et al. (1997). "Clathrin-mediated Endocytosis of the -Adrenergic Receptor Is Regulated by Phosphorylation/Dephosphorylation of -Arrestin1". In: *The Journal of biological chemistry* 272.49, pp. 31051–31057. ISSN: 0021-9258. DOI: 10.1074/jbc.272.49.31051.

Lin, F.-T. et al. (1999). "Feedback Regulation of β-Arrestin1 Function by Extracellular Signal-regulated Kinases". In: *The Journal of biological chemistry* 274.23, pp. 15971–15974. ISSN: 0021-9258. DOI: 10.1074/jbc.274.23.15971.

Lin, F.-T. et al. (2002). "Phosphorylation of β-Arrestin2 Regulates Its Function in Internalization of β 2 -Adrenergic Receptors". In: *Biochemistry* 41.34, pp. 10692–10699. ISSN: 0006-2960. DOI: 10.1021/bi025705n.

Lin, Y.-S. et al. (2007). "Proportion of solvent-exposed amino acids in a protein and rate of protein evolution". In: *Molecular biology and evolution* 24.4, pp. 1005–1011. ISSN: 0737-4038. DOI: 10.1093/molbev/msm019.

Lin, Z. and W.-H. Li (2011). "Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts". In: *Molecular biology and evolution* 28.1, pp. 131–142. ISSN: 0737-4038. DOI: 10.1093/molbev/msq184.

Lisney, T. J. et al. (2012). "Vision in elasmobranchs and their relatives: 21st century advances". In: *Journal of fish biology* 80.5, pp. 2024–2054. ISSN: 0022-1112. DOI: 10.1111/j.1095-8649.2012.03253.x.

Liu, Y. et al. (2015). "Biased signalling: the instinctive skill of the cell in the selection of appropriate signalling pathways". In: *The Biochemical journal* 470.2, pp. 155–167. ISSN: 1470-8728. DOI: 10.1042/BJ20150358.

Lohse, M. et al. (1990). "beta-Arrestin: A protein that regulates beta-adrenergic receptor function". In: *Science (New York, N.Y.)* 248.4962, pp. 1547–1550. ISSN: 1095-9203. DOI: 10.1126/science.2163110.

Lokits, A. D. et al. (2018). "Tracing the evolution of the heterotrimeric G protein α subunit in Metazoa: In revision". In: *BMC evolutionary biology*. ISSN: 1471-2148.

Lovász, L. and M. D. Plummer (1986). *Matching Theory*. Vol. 29. Amsterdam, NL: Elsevier.

Lovgren, A. K. et al. (2011). "β-arrestin deficiency protects against pulmonary fibrosis in mice and prevents fibroblast invasion of extracellular matrix". In: *Science translational medicine* 3.74, 74ra23. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.3001564.

Lozada-Chávez, I., P. F. Stadler, and S. J. Prohaska (2018). "Genome-wide features of introns are evolutionary decoupled among themselves and from genome size throughout Eukarya: In revision". In:

Luan, B. et al. (2009). "Deficiency of a beta-arrestin-2 signal complex contributes to insulin resistance". In: *Nature* 457.7233, pp. 1146–1149. ISSN: 1476-4687. DOI: 10.1038/nature07617.

Lunzer, M. et al. (2010). "Correction: Pervasive Cryptic Epistasis in Molecular Evolution". In: *PLoS Genetics* 6.11. ISSN: 1553-7404. DOI: 10.1371/annotation/d618ce28-5010-47df-a44e-148ecdc0fef6.

Luttrell, L. M. (1999). "Beta-Arrestin-Dependent Formation of 2 Adrenergic Receptor-Src Protein Kinase Complexes". In: *Science (New York, N.Y.)* 283.5402, pp. 655–661. ISSN: 1095-9203. DOI: 10.1126/science.283.5402.655.

– (2013). *Molecular biology of arrestins*. Vol. v. 118. Progress in molecular biology and translational science. Oxford: Elsevier Science & Technology. ISBN: 9781299666481.

Lykke-Andersen, S. and T. H. Jensen (2015). "Nonsense-mediated mRNA decay: An intricate machinery that shapes transcriptomes". In: *Nature reviews. Molecular cell biology* 16.11, pp. 665–677. ISSN: 1471-0072. DOI: 10.1038/nrm4063.

Lymperopoulos, A. and A. Bathgate (2013). "Arrestins in the cardiovascular system". In: *Progress in molecular biology and translational science* 118, pp. 297–334. ISSN: 1878-0814. DOI: 10.1016/B978-0-12-394440-5.00012-7.

Ma, X. et al. (2012). "Acute activation of β2-adrenergic receptor regulates focal adhesions through βArrestin2- and p115RhoGEF protein-mediated activation of RhoA". In: *The Journal of biological chemistry* 287.23, pp. 18925–18936. ISSN: 0021-9258. DOI: 10.1074/jbc.M112.352260.

Maeda, T. et al. (2000). "Purification and characterization of bovine cone arrestin (cArr)". In: *FEBS letters* 470.3, pp. 336–340. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(00)01334-X.

Magadum, S. et al. (2013). "Gene duplication as a major force in evolution". In: *Journal of genetics* 92.1, pp. 155–161. ISSN: 0022-1333. DOI: 10.1007/s12041-013-0212-8.

Mahabaleshwar, H. et al. (2012). "beta-arrestin control of late endosomal sorting facilitates decoy receptor function and chemokine gradient formation". In: *Development (Cambridge, England)* 139.16, pp. 2897–2902. ISSN: 1477-9129. DOI: 10.1242/dev.080408.

Mani, S. S., J. C. Besharse, and B. E. Knox (1999). "Immediate upstream sequence of arrestin directs rod-specific expression in Xenopus". In: *The Journal of biological chemistry* 274.22, pp. 15590–15597. ISSN: 0021-9258. DOI: 10.1074/jbc.274.22.15590.

Manning, J. R., E. R. Jefferson, and G. J. Barton (2008). "The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction". In: *BMC bioinformatics* 9, p. 51. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-51.

Manuel, M. de et al. (2016). "Chimpanzee genomic diversity reveals ancient admixture with bonobos". In: *Science (New York, N.Y.)* 354.6311, pp. 477–481. ISSN: 1095-9203. DOI: 10.1126/science.aag2602.

Marion, S. et al. (2007). "N-terminal tyrosine modulation of the endocytic adaptor function of the beta-arrestins". In: *The Journal of biological chemistry* 282.26, pp. 18937–18944. ISSN: 0021-9258. DOI: 10.1074/jbc.M700090200.

Maronde, E. et al. (2011). "Dynamics in enzymatic protein complexes offer a novel principle for the regulation of melatonin synthesis in the human pineal gland". In: *Journal of pineal research* 51.1, pp. 145–155. ISSN: 0742-3098. DOI: 10.1111/j.1600-079X.2011.00880.x.

Marsh, J. A. and S. A. Teichmann (2010). "How do proteins gain new domains?" In: *Genome biology* 11.7, p. 126. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-7-126.

Martin, D. P. et al. (2015). "RDP4: Detection and analysis of recombination patterns in virus genomes". In: *Virus evolution* 1.1, vev003. ISSN: 2057-1577. DOI: 10.1093/ve/vev003.

Masuda, S. et al. (2005). "Recruitment of the human TREX complex to mRNA during splicing". In: *Genes & development* 19.13, pp. 1512–1517. ISSN: 0890-9369. DOI: 10.1101/gad.1302205.

Materna, S. C., K. Berney, and R. A. Cameron (2006). "The S. purpuratus genome: a comparative perspective". In: *Developmental biology* 300.1, pp. 485–495. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2006.09.033.

Mayeenuddin, L. H. and J. Mitchell (2003). "Squid visual arrestin: CDNA cloning and calcium-dependent phosphorylation by rhodopsin kinase (SQRK)". In: *Journal of neurochemistry* 85.3, pp. 592–600. ISSN: 0022-3042. DOI: 10.1046/j.1471-4159.2003.01726.x.

McDonald, P. H. et al. (2000). "Beta-arrestin 2: a receptor-regulated MAPK scaffold for the activation of JNK3". In: *Science (New York, N.Y.)* 290.5496, pp. 1574–1577. ISSN: 1095-9203. DOI: 10.1126/science.290.5496.1574.

McQuilton, P. et al. (2012). "FlyBase 101 – the basics of navigating FlyBase". In: *Nucleic acids research* 40/D1. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1030.

Mehta, T. K. et al. (2013). "Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum)". In: *Proceedings of the National Academy of Sciences* 110.40, pp. 16044–16049. ISSN: 1091-6490. DOI: 10.1073/pnas.1315760110.

Mendoza, A. de, A. Sebé-Pedrós, and I. Ruiz-Trillo (2014). "The evolution of the GPCR signaling system in eukaryotes: modularity, conservation, and the transition to metazoan multicellularity". In: *Genome biology and evolution* 6.3, pp. 606–619. ISSN: 1759-6653. DOI: 10.1093/gbe/evu038.

Mertins, P. et al. (2013). "Integrated proteomic analysis of post-translational modifications by serial enrichment". In: *Nature methods* 10.7, pp. 634–637. ISSN: 1548-7105. DOI: 10.1038/nmeth.2518.

Mertins, P. et al. (2016). "Proteogenomics connects somatic mutations to signalling in breast cancer". In: *Nature* 534.7605, pp. 55–62. ISSN: 1476-4687. DOI: 10.1038/nature18003.

Metropolis, N. et al. (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 0021-9606. DOI: 10.1063/1.1699114.

Meyer, A. and Y. van de Peer (2005). "From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD)". In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 27.9, pp. 937–945. ISSN: 0265-9247. DOI: 10.1002/bies.20293.

Mikkelsen, T. S. et al. (2007). "Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences". In: *Nature* 447.7141, pp. 167–177. ISSN: 1476-4687. DOI: 10.1038/nature05805.

Milano, S. K. et al. (2002). "Scaffolding Functions of Arrestin-2 Revealed by Crystal Structure and Mutagenesis". In: *Biochemistry* 41.10, pp. 3321–3328. ISSN: 0006-2960. DOI: 10.1021/bi015905j.

Milano, S. K. et al. (2006). "Nonvisual arrestin oligomerization and cellular localization are regulated by inositol hexakisphosphate binding". In: *The Journal of biological chemistry* 281.14, pp. 9812–9823. ISSN: 0021-9258. DOI: 10.1074/jbc.M512703200.

Minning, J., M. Porto, and U. Bastolla (2013). "Detecting selection for negative design in proteins through an improved model of the misfolded state". In: *Proteins* 81.7, pp. 1102–1112. ISSN: 0887-3585. DOI: 10.1002/prot.24244.

Moaven, H. et al. (2013). "Visual arrestin interaction with clathrin adaptor AP-2 regulates photoreceptor survival in the vertebrate retina". In: *Proceedings of the National Academy of Sciences* 110.23, pp. 9463–9468. ISSN: 1091-6490. DOI: 10.1073/pnas.1301126110.

Molnar, C. et al. (2011). "Role of the Drosophila non-visual ß-arrestin kurtz in hedgehog signalling". In: *PLoS Genetics* 7.3, e1001335. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001335.

Morshedian, A. and G. L. Fain (2015). "Single-photon sensitivity of lamprey rods with cone-like outer segments". In: *Current Biology* 25.4, pp. 484–487. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.12.031.

Mosser, V. A. et al. (2008). "Differential role of beta-arrestin ubiquitination in agonist-promoted down-regulation of M1 vs M2 muscarinic acetylcholine receptors". In: *Journal of molecular signaling* 3, p. 20. ISSN: 1750-2187. DOI: 10.1186/1750-2187-3-20.

Mukherjee, A. et al. (2005). "Regulation of Notch signalling by non-visual beta-arrestin". In: *Nature cell biology* 7.12, pp. 1191–1201. ISSN: 1465-7392. DOI: 10.1038/ncb1327.

Nakagawa, M. et al. (2002). "Ascidian arrestin (Ci-arr), the origin of the visual and nonvisual arrestins of vertebrate". In: *European journal of biochemistry / FEBS* 269.21, pp. 5112–5118. ISSN: 0014-2956. DOI: 10.1046/j.1432-1033.2002.03240.x.

Nalivaeva, N. N. and A. J. Turner (2009). "Lipid Anchors to Proteins". In: *Handbook of Neurochemistry and Molecular Neurobiology*. Ed. by A. Lajtha, G. Goracci, and G. Tettamanti. Boston, MA: Springer, pp. 353–372. ISBN: 978-0-387-30345-1. DOI: 10.1007/978-0-387-30378-9_14.

National Center for Biotechnology Information (2017). *Genome Report*. URL: ftp://ftp.ncbi.nih.gov/genomes/GENOME_REPORTS/eukaryotes.txt.

National Institute of Health (US) and National Center for Biotechnology Information. *Translated BLAST: tblastn*. Ed. by National Institute of Health (US) and National

Center for Biotechnology Information. URL: http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=tblastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome (visited on 11/09/2016).

Neuhaus, E. M. et al. (2006). "Novel function of beta-arrestin2 in the nucleus of mature spermatozoa". In: *Journal of cell science* 119.Pt 15, pp. 3047–3056. ISSN: 0021-9533. DOI: 10.1242/jcs.03046.

Nikonov, S. S. et al. (2008). "Mouse cones require an arrestin for normal inactivation of phototransduction". In: *Neuron* 59.3, pp. 462–474. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.06.011.

Nilsson, M. A. et al. (2010). "Tracking marsupial evolution using archaic genomic retroposon insertions". In: *PLoS biology* 8.7, e1000436. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000436.

Nilsson, M. A. et al. (2012). "Expansion of CORE-SINEs in the genome of the Tasmanian devil". In: *BMC genomics* 13, p. 172. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-172.

Nitsche, A. et al. (2015). "Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved". In: *RNA (New York, N.Y.)* 21.5, pp. 801–812. ISSN: 1469-9001. DOI: 10.1261/rna.046342.114.

Noivirt-Brik, O., A. Horovitz, and R. Unger (2009). "Trade-off between positive and negative design of protein stability: from lattice models to real proteins". In: *PLoS computational biology* 5.12, e1000592. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000592.

Nordström, K., T. A. Larsson, and D. Larhammar (2004). "Extensive duplications of phototransduction genes in early vertebrate evolution correlate with block (chromosome) duplications". In: *Genomics* 83.5, pp. 852–872. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2003.11.008.

Nowick, K. et al. (2011). "Gain, Loss and Divergence in Primate Zinc-Finger Genes: A Rich Resource for Evolution of Gene Regulatory Differences between Species". In: *PloS one* 6.6. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021553.

Oakley, R. H. et al. (2000). "Differential affinities of visual arrestin, beta arrestin1, and beta arrestin2 for G protein-coupled receptors delineate two major classes of receptors". In: *The Journal of biological chemistry* 275.22, pp. 17201–17210. ISSN: 0021-9258. DOI: 10.1074/jbc.M910348199.

O'Brien, K. P., M. Remm, and E. L. L. Sonnhammer (2005). "Inparanoid: a comprehensive database of eukaryotic orthologs". In: *Nucleic acids research* 33, pp. D476–80. ISSN: 1362-4962. DOI: 10.1093/nar/gki107.

Ohno, S. (1970). *Evolution by gene duplication*. Berlin and New York: Springer. ISBN: 3642866611.

O'Leary, N. A. et al. (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic acids research* 44.D1, pp. D733–45. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

OpenStax CNX, ed. (2017-09-13). *Biology*. URL: http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.118 (visited on 11/14/2017).

Orem, N. R. and P. J. Dolph (2002). "Epitope masking of rhabdomeric rhodopsin during endocytosis-induced retinal degeneration". In: *Molecular vision* 8, pp. 455–461. ISSN: 1090-0535.

Orquera, D. P. and F. S. J. de Souza (2017). "Evolution of the Rax family of developmental transcription factors in vertebrates". In: *Mechanisms of development* 144.Pt B, pp. 163–170. ISSN: 1872-6356. DOI: 10.1016/j.mod.2016.11.002.

Ostermaier, M. K. et al. (2014). "Functional map of arrestin-1 at single amino acid resolution". In: *Proceedings of the National Academy of Sciences* 111.5, pp. 1825–1830. ISSN: 1091-6490. DOI: 10.1073/pnas.1319402111.

Ozawa, K. et al. (2008). "S-nitrosylation of beta-arrestin regulates beta-adrenergic receptor trafficking". In: *Molecular cell* 31.3, pp. 395–405. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2008.05.024.

Pace, C. N. and J. Hermans (2008). "The Stability of Globular Protein". In: *CRC Critical Reviews in Biochemistry* 3.1, pp. 1–43. ISSN: 0045-6411. DOI: 10.3109/10409237509102551.

Page, R. D. M. and E. C. Holmes (2005). *Molecular evolution: A phylogenetic approach*. 9. print. Malden Mass a.o.: Wiley-Blackwell. ISBN: 978-0-86542-889-8.

Pain, D. et al. (2005). "Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification". In: *The Journal of biological chemistry* 280.8, pp. 6265–6268. ISSN: 0021-9258. DOI: 10.1074/jbc.C400587200.

Palczewski, K. and W. C. Smith (1996). "Splice variants of arrestins". In: *Experimental eye research* 63.5, pp. 599–602. ISSN: 0014-4835. DOI: 10.1006/exer.1996.0151.

Parada, G. E. et al. (2014). "A comprehensive survey of non-canonical splice sites in the human transcriptome". In: *Nucleic acids research* 42.16, pp. 10564–10578. ISSN: 1362-4962. DOI: 10.1093/nar/gku744.

Paradis, J. S. et al. (2015). "Receptor sequestration in response to β-arrestin-2 phosphorylation by ERK1/2 governs steady-state levels of GPCR cell-surface expression". In: *Proceedings of the National Academy of Sciences* 112.37, E5160–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1508836112.

Parathath, S. R. et al. (2010). "β-Arrestin-1 links mitogenic sonic hedgehog signaling to the cell cycle exit machinery in neural precursors". In: *Cell cycle (Georgetown, Tex.)* 9.19, pp. 4013–4024. ISSN: 1551-4005. DOI: 10.4161/cc.9.19.13325.

Parra, G. et al. (2003). "Comparative gene prediction in human and mouse". In: *Genome research* 13.1, pp. 108–117. ISSN: 1088-9051. DOI: 10.1101/gr.871403.

Parruti, G. et al. (1993). "Molecular analysis of human beta-arrestin-1: cloning, tissue distribution, and regulation of expression. Identification of two isoforms generated by alternative splicing". In: *The Journal of biological chemistry* 268.13, pp. 9753–9761. ISSN: 0021-9258.

Pauling, L. and R. B. Corey (1951). "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets". In: *Proceedings of the National Academy of Sciences* 37.11, pp. 729–740. ISSN: 1091-6490. DOI: 10.1073/pnas.37.11.729.

Pauling, L., R. B. Corey, and H. R. Branson (1951). "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain". In: *Proceedings of the National Academy of Sciences* 37.4, pp. 205–211. ISSN: 1091-6490. DOI: 10.1073/pnas.37.4.205.

Pavesi, G. et al. (2008). "Exalign: a new method for comparative analysis of exon-intron gene structures". In: *Nucleic acids research* 36.8, e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkn153.

Pazos, F., A. Rausell, and A. Valencia (2006). "Phylogeny-independent detection of functional residues". In: *Bioinformatics* 22.12, pp. 1440–1448. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl104.

Peiro, G. et al. (2004). "Analysis of HER-2/neu amplification in endometrial carcinoma by chromogenic in situ hybridization. Correlation with fluorescence in situ hybridization, HER-2/neu, p53 and Ki-67 protein expression, and outcome". In:

Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 17.3, pp. 227–287. ISSN: 0893-3952. DOI: 10.1038/modpathol.3800006.

Petryszak, R. et al. (2016). "Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants". In: *Nucleic acids research* 44.D1, pp. D746–52. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1045.

Pickrell, S. W. et al. (2004). "Deciphering the contribution of known cis-elements in the mouse cone arrestin gene to its cone-specific expression". In: *Investigative ophthalmology & visual science* 45.11, pp. 3877–3884. ISSN: 0146-0404. DOI: 10.1167/iovs.04-0663.

Pillmann, H. et al. (2011). "Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology". In: *BMC bioinformatics* 12.1, p. 270. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-270.

Piper, R. C., I. Dikic, and G. L. Lukacs (2014). "Ubiquitin-dependent sorting in endocytosis". In: *Cold Spring Harbor perspectives in biology* 6.1. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a016808.

Pirovano, W., K. A. Feenstra, and J. Heringa (2006). "Sequence comparison by sequence harmony identifies subtype-specific functional sites". In: *Nucleic acids research* 34.22, pp. 6540–6548. ISSN: 1362-4962. DOI: 10.1093/nar/gkl901.

Plachetzki, D. C., C. R. Fong, and T. H. Oakley (2012). "Cnidocyte discharge is regulated by light and opsin-mediated phototransduction". In: *BMC biology* 10, p. 17. ISSN: 1741-7007. DOI: 10.1186/1741-7007-10-17.

Polekhina, G. et al. (2013). "Structure of the N-terminal domain of human thioredoxin-interacting protein". In: *Acta crystallographica. Section D, Biological crystallography* 69.Pt 3, pp. 333–344. ISSN: 1399-0047. DOI: 10.1107/S0907444912047099.

Por, E. D. et al. (2013). "Phosphorylation regulates TRPV1 association with β-arrestin-2". In: *The Biochemical journal* 451.1, pp. 101–109. ISSN: 1470-8728. DOI: 10.1042/BJ20121637.

Posada, D. (2008). "jModelTest: phylogenetic model averaging". In: *Molecular biology and evolution* 25.7, pp. 1253–1256. ISSN: 0737-4038. DOI: 10.1093/molbev/msn083.

Potrzebowski, L. et al. (2008). "Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes". In: *PLoS biology* 6.4, e80. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060080.

Price, M. N., P. S. Dehal, and A. P. Arkin (2010). "FastTree 2–approximately maximum-likelihood trees for large alignments". In: *PloS one* 5.3, e9490. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490.

Puca, L. et al. (2013). "α-arrestin 1 (ARRDC1) and β-arrestins cooperate to mediate Notch degradation in mammals". In: *Journal of cell science* 126.Pt 19, pp. 4457–4468. ISSN: 0021-9533. DOI: 10.1242/jcs.130500.

Purayil, H. T. et al. (2015). "Arrestin2 modulates androgen receptor activation". In: *Oncogene* 34.24, pp. 3144–3151. ISSN: 1476-5594. DOI: 10.1038/onc.2014.252.

Quax, T. E. F. et al. (2015). "Codon Bias as a Means to Fine-Tune Gene Expression". In: *Molecular cell* 59.2, pp. 149–161. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.05.035.

Ragg, H. et al. (2009). "Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes". In: *BMC evolutionary biology* 9, p. 208. ISSN: 1471-2148. DOI: 10.1186/1471-2148-9-208.

Raible, F. et al. (2006). "Opsins and clusters of sensory G-protein-coupled receptors in the sea urchin genome". In: *Developmental biology* 300.1, pp. 461–475. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2006.08.070.

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan (1963). "Stereochemistry of polypeptide chain configurations". In: *Journal of molecular biology* 7.1, pp. 95–99. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(63)80023-6.

Rambaut, A., M. A. Suchard, and A. J. Drummond (2014). *Tracer*. URL: http://beast.bio.ed.ac.uk/Tracer (visited on 04/24/2017).

Ranwez, V. et al. (2011). "MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons". In: *PloS one* 6.9, e22594. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022594.

Raumbaut, A. (2006). *FigTree: Tree Figure Drawing Tool*. URL: http://tree.bio.ed.ac.uk/ (visited on 05/20/2017).

Rausell, A. et al. (2010). "Protein interactions and ligand binding: from protein subfamilies to functional specificity". In: *Proceedings of the National Academy of Sciences* 107.5, pp. 1995–2000. ISSN: 1091-6490. DOI: 10.1073/pnas.0908044107.

Ren, X.-R. et al. (2005). "Different G protein-coupled receptor kinases govern G protein and beta-arrestin-mediated signaling of V2 vasopressin receptor". In: *Proceedings of the National Academy of Sciences* 102.5, pp. 1448–1453. ISSN: 1091-6490. DOI: 10.1073/pnas.0409534102.

Renninger, S. L., M. Gesemann, and S. C. F. Neuhauss (2011). "Cone arrestin confers cone vision of high temporal resolution in zebrafish larvae". In: *The European journal of neuroscience* 33.4, pp. 658–667. ISSN: 0953-816X. DOI: 10.1111/j.1460-9568.2010.07574.x.

Replogle, K. et al. (2008). "The Songbird Neurogenomics (SoNG) Initiative: community-based tools and strategies for study of brain gene function and evolution". In: *BMC genomics* 9, p. 131. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-131.

Rios, S. et al. (2015). "GPCRtm: An amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors". In: *BMC bioinformatics* 16, p. 206. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0639-4.

Robles, M. S., S. J. Humphrey, and M. Mann (2017). "Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology". In: *Cell metabolism* 25.1, pp. 118–127. ISSN: 1932-7420. DOI: 10.1016/j.cmet.2016.10.004.

Rogozin, I. B. et al. (2012). "Origin and evolution of spliceosomal introns". In: *Biology direct* 7, p. 11. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-11.

Roux, J., J. Liu, and M. Robinson-Rechavi (2017). "Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates". In: *Molecular biology and evolution*. ISSN: 0737-4038. DOI: 10.1093/molbev/msx199.

Roy, S. W. and D. Penny (2007). "On the incidence of intron loss and gain in paralogous gene families". In: *Molecular biology and evolution* 24.8, pp. 1579–1581. ISSN: 0737-4038. DOI: 10.1093/molbev/msm082.

Samuel, A. et al. (2014). "Otx2 ChIP-seq reveals unique and redundant functions in the mature mouse retina". In: *PloS one* 9.2, e89110. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0089110.

Santos-Zas, I. et al. (2013). "β-Arrestin signal complex plays a critical role in adipose differentiation". In: *The international journal of biochemistry & cell biology* 45.7, pp. 1281–1292. ISSN: 1878-5875. DOI: 10.1016/j.biocel.2013.03.014.

Sapède, D. and E. Cau (2013). "The pineal gland from development to function". In: *Current topics in developmental biology* 106, pp. 171–215. ISSN: 1557-8933. DOI: 10.1016/B978-0-12-416021-7.00005-5.

Sato, K. et al. (2011). "Vertebrate ancient-long opsin has molecular properties intermediate between those of vertebrate and invertebrate visual pigments". In: *Biochemistry* 50.48, pp. 10484–10490. ISSN: 0006-2960. DOI: 10.1021/bi201212z.

Sato, Y., Y. Hashiguchi, and M. Nishida (2009). "Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication". In: *BMC evolutionary biology* 9, p. 127. ISSN: 1471-2148. DOI: 10.1186/1471-2148-9-127.

Satoh, A. K. and D. F. Ready (2005). "Arrestin1 mediates light-dependent rhodopsin endocytosis and cell survival". In: *Current Biology* 15.19, pp. 1722–1733. ISSN: 0960-9822. DOI: 10.1016/j.cub.2005.08.064.

Sayers, E. W. et al. (2012). "Database resources of the National Center for Biotechnology Information". In: *Nucleic acids research* 40, pp. D13–25. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1184.

Scherer, S. (2010). *Guide to the human genome*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. ISBN: 0879699442.

Schmelzle, K. and F. M. White (2006). "Phosphoproteomic approaches to elucidate cellular signaling networks". In: *Current opinion in biotechnology* 17.4, pp. 406–414. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2006.06.004.

Schmid, E. M. et al. (2006). "Role of the AP2 beta-appendage hub in recruiting partners for clathrin-coated vesicle assembly". In: *PLoS biology* 4.9, e262. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040262.

Schrödinger, L. (2015). *The PyMOL Molecular Graphics System*.

Schulte, G., A. Schambony, and V. Bryja (2010). "beta-Arrestins - scaffolds and signalling elements essential for WNT/Frizzled signalling pathways?" In: *British journal of pharmacology* 159.5, pp. 1051–1058. ISSN: 1476-5381. DOI: 10.1111/j.1476-5381.2009.00466.x.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464. ISSN: 0090-5364. DOI: 10.1214/aos/1176344136.

Scott, M. G. H. et al. (2002). "Differential Nucleocytoplasmic Shuttling of β-Arrestins CHARACTERIZATION OF A LEUCINE-RICH NUCLEAR EXPORT SIGNAL IN β-ARRESTIN2". In: *The Journal of biological chemistry* 277.40, pp. 37693–37701. ISSN: 0021-9258. DOI: 10.1074/jbc.M207552200.

Seo, J. et al. (2011). "Identification of arrestin-3-specific residues necessary for JNK3 kinase activation". In: *The Journal of biological chemistry* 286.32, pp. 27894–27901. ISSN: 0021-9258. DOI: 10.1074/jbc.M111.260448.

Setta, N. de et al. (2014). "Building the sugarcane genome for biotechnology and identifying evolutionary trends". In: *BMC genomics* 15, p. 540. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-540.

Shankar, H. et al. (2010). "Non-visual Arrestins Are Constitutively Associated with the Centrosome and Regulate Centrosome Function*". In: *The Journal of biological chemistry* 285.11, pp. 8316–8329. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.062521.

Sharma, V., P. Schwede, and M. Hiller (2017). "CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation". In: *Bioinformatics*. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx527.

Shenoy, S. K. and R. J. Lefkowitz (2003). "Trafficking Patterns of β-Arrestin and G Protein-coupled Receptors Determined by the Kinetics of β-Arrestin Deubiquitination". In: *The Journal of biological chemistry* 278.16, pp. 14498–14506. ISSN: 0021-9258. DOI: 10.1074/jbc.M209626200.

– (2005). "Receptor-specific ubiquitination of beta-arrestin directs assembly and targeting of seven-transmembrane receptor signalosomes". In: *The Journal of biological chemistry* 280.15, pp. 15315–15324. ISSN: 0021-9258. DOI: 10.1074/jbc.M412418200.
– (2011). "β-Arrestin-mediated receptor trafficking and signal transduction". In: *Trends in pharmacological sciences* 32.9, pp. 521–533. ISSN: 1873-3735. DOI: 10.1016/j.tips.2011.05.002.
Shenoy, S. K. et al. (2001). "Regulation of receptor fate by ubiquitination of activated beta 2-adrenergic receptor and beta-arrestin". In: *Science (New York, N.Y.)* 294.5545, pp. 1307–1313. ISSN: 1095-9203. DOI: 10.1126/science.1063866.
Shenoy, S. K. et al. (2009). "Beta-arrestin-dependent signaling and trafficking of 7-transmembrane receptors is reciprocally regulated by the deubiquitinase USP33 and the E3 ligase Mdm2". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6650–6655. ISSN: 1091-6490. DOI: 10.1073/pnas.0901083106.
Shepelev, V. and A. Fedorov (2006). "Advances in the Exon-Intron Database (EID)". In: *Briefings in bioinformatics* 7.2, pp. 178–185. ISSN: 1467-5463. DOI: 10.1093/bib/bbl003.
Shi, G. et al. (2007). "Signaling properties of a short-wave cone visual pigment and its role in phototransduction". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 27.38, pp. 10084–10093. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2211-07.2007.
Shi, H. et al. (2006). "The retromer subunit Vps26 has an arrestin fold and binds Vps35 through its C-terminal domain". In: *Nature structural & molecular biology* 13.6, pp. 540–548. ISSN: 1545-9993. DOI: 10.1038/nsmb1103.
Shukla, A. K. et al. (2013). "Structure of active β-arrestin-1 bound to a G-protein-coupled receptor phosphopeptide". In: *Nature* 497.7447, pp. 137–141. ISSN: 1476-4687. DOI: 10.1038/nature12120.
Sibley, C. R., J. Ule, and L. Blazquez (2016). "Lessons from non-canonical splicing". In: *Nature reviews. Genetics* 17.7, p. 407. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.46.
Siepel, A. (2009). "Phylogenomics of primates and their ancestral populations". In: *Genome research* 19.11, pp. 1929–1941. ISSN: 1088-9051. DOI: 10.1101/gr.084228.108.
Sievers, F. et al. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular systems biology* 7, p. 539. ISSN: 1744-4292. DOI: 10.1038/msb.2011.75.
Silva, J.-P. and Y. A. Ushkaryov (2011). "The Latrophilins, "Split-Personality" Receptors". In: *Adhesion-GPCRs*. Ed. by S. Yona and M. Stacey. Vol. 706. Advances in Experimental Medicine and Biology. Boston, MA: Landes Bioscience and Springer Science+Business Media LLC, pp. 59–75. ISBN: 978-1-4419-7913-1. DOI: 10.1007/978-1-4419-7913-1_5.
Singh, P. P. et al. (2014). "Human dominant disease genes are enriched in paralogs originating from whole genome duplication". In: *PLoS computational biology* 10.7, e1003754. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003754.
Slater, G. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison". In: *BMC bioinformatics* 6, p. 31. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-31.
Smith, J. J. and M. C. Keinath (2015). "The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications". In: *Genome research* 25.8, pp. 1081–1090. ISSN: 1088-9051. DOI: 10.1101/gr.184135.114.

Smith, J. J. et al. (2012). "Genetic consequences of programmed genome rearrangement". In: *Current Biology* 22.16, pp. 1524–1529. ISSN: 0960-9822. DOI: 10.1016/j.cub.2012.06.028.

Smith, J. J. et al. (2013). "Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution". In: *Nature genetics* 45.4, 415–21, 421e1–2. ISSN: 1061-4036. DOI: 10.1038/ng.2568.

Smith, W. C. (1996). "A splice variant of arrestin from human retina". In: *Experimental eye research* 62.6, pp. 585–592. ISSN: 0014-4835.

Smith, W. C. et al. (1994). "A splice variant of arrestin. Molecular cloning and localization in bovine retina". In: *The Journal of biological chemistry* 269.22, pp. 15407–15410. ISSN: 0021-9258.

Smith, W. C. et al. (2000). "Cloning and functional characterization of salamander rod and cone arrestins". In: *Investigative ophthalmology & visual science* 41.9, pp. 2445–2455. ISSN: 0146-0404.

Smith, W. C. et al. (1995). "Isolation and Expression of an Arrestin cDNA from the Horseshoe Crab Lateral Eye". In: *Journal of neurochemistry* 64.1, pp. 1–13. ISSN: 0022-3042. DOI: 10.1046/j.1471-4159.1995.64010001.x.

Softberry, I. (2007). *Prot_map*. URL: http://linux1.softberry.com/berry.phtml?topic=prot_map&group=help&subgroup=xmap.

Solessio, E. and G. A. Engbretson (1993). "Antagonistic chromatic mechanisms in photoreceptors of the parietal eye of lizards". In: *Nature* 364.6436, pp. 442–445. ISSN: 1476-4687. DOI: 10.1038/364442a0.

Soltis, P. S. and D. E. Soltis, eds. (2012). *Polyploidy and genome evolution*. Berlin and New York: Springer. ISBN: 9783642432811.

Song, X., E. V. Gurevich, and V. V. Gurevich (2007). "Cone arrestin binding to JNK3 and Mdm2: conformational preference and localization of interaction sites". In: *Journal of neurochemistry* 103.3, pp. 1053–1062. ISSN: 0022-3042. DOI: 10.1111/j.1471-4159.2007.04842.x.

Song, X. et al. (2009). "How does arrestin assemble MAPKs into a signaling complex?" In: *The Journal of biological chemistry* 284.1, pp. 685–695. ISSN: 0021-9258. DOI: 10.1074/jbc.M806124200.

Song, X. et al. (2011). "Arrestin-1 expression level in rods: balancing functional performance and photoreceptor health". In: *Neuroscience* 174, pp. 37–49. ISSN: 0306-4522. DOI: 10.1016/j.neuroscience.2010.11.009.

Sonoda, Y., K. Mizutani, and B. Mikami (2015). "Structure of Spo0M, a sporulation-control protein from Bacillus subtilis". In: *Acta crystallographica. Section F, Structural biology communications* 71.Pt 12, pp. 1488–1497. ISSN: 2053-230X. DOI: 10.1107/S2053230X15020919.

Stamatakis, A. (2006). "Phylogenetic models of rate heterogeneity: A high performance computing perspective". In: *Proceedings of the 20th international conference on Parallel and distributed processing*. Washington, DC: IEEE Computer Society, 8 pp. ISBN: 1-4244-0054-6. DOI: 10.1109/IPDPS.2006.1639535.

Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel". In: *Bioinformatics* 19.Suppl 2, pp. ii215–ii225. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg1080.

Stergachis, A. B. et al. (2013). "Exonic transcription factor binding directs codon choice and affects protein evolution". In: *Science (New York, N.Y.)* 342.6164, pp. 1367–1372. ISSN: 1095-9203. DOI: 10.1126/science.1243490.

Sterne-Marr, R. et al. (1993). "Polypeptide variants of beta-arrestin and arrestin3". In: *The Journal of biological chemistry* 268.21, pp. 15640–15648. ISSN: 0021-9258.

Storto, M. (2001). "Expression of metabotropic glutamate receptors in the rat and human testis". In: *Journal of Endocrinology* 170.1, pp. 71–78. ISSN: 0022-0795. DOI: 10.1677/joe.0.1700071.

Stover, B. C. and K. F. Muller (2010). "TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses". In: *BMC bioinformatics* 11, p. 7. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-7.

Strungs, E. G. and L. M. Luttrell (2014). "Arrestin-dependent activation of ERK and Src family kinases". In: *Handbook of experimental pharmacology* 219, pp. 225–257. ISSN: 0171-2004. DOI: 10.1007/978-3-642-41199-1_12.

Suarez, H. G. et al. (2017). "chainCleaner improves genome alignment specificity and sensitivity". In: *Bioinformatics* 33.11, pp. 1596–1603. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx024.

Sun, J.-J. et al. (2016). "β-Arrestin 1's Interaction with TC45 Attenuates Stat signaling by dephosphorylating Stat to inhibit antimicrobial peptide expression". In: *Scientific Reports* 6, p. 35808. ISSN: 2045-2322. DOI: 10.1038/srep35808.

Sutton, R. B. et al. (2005). "Crystal structure of cone arrestin at 2.3A: Evolution of receptor specificity". In: *Journal of molecular biology* 354.5, pp. 1069–1080. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2005.10.023.

Sverdlov, A. V. et al. (2004). "Reconstruction of ancestral protosplice sites". In: *Current Biology* 14.16, pp. 1505–1508. ISSN: 0960-9822. DOI: 10.1016/j.cub.2004.08.027.

Szczepek, M. et al. (2014). "Crystal structure of a common GPCR-binding interface for G protein and arrestin". In: *Nature communications* 5, p. 4801. ISSN: 2041-1723. DOI: 10.1038/ncomms5801.

Tamura, K. et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0". In: *Molecular biology and evolution* 24.8, pp. 1596–1599. ISSN: 0737-4038. DOI: 10.1093/molbev/msm092.

ter Haar, E., S. C. Harrison, and T. Kirchhausen (2000). "Peptide-in-groove interactions link target proteins to the beta-propeller of clathrin". In: *Proceedings of the National Academy of Sciences* 97.3, pp. 1096–1100. ISSN: 1091-6490.

Terakita, A. (2005). "The opsins". In: *Genome biology* 6.3, p. 213. ISSN: 1465-6906. DOI: 10.1186/gb-2005-6-3-213.

Terakita, A., E. Kawano-Yamashita, and M. Koyanagi (2012). "Evolution and diversity of opsins". In: *Wiley interdisciplinary reviews. Membrane transport and signaling* 1.1, pp. 104–111. ISSN: 2190-460X. DOI: 10.1002/wmts.6.

Thathiah, A. and B. de Strooper (2011). "The role of G protein-coupled receptors in the pathology of Alzheimer's disease". In: *Nature Reviews. Neuroscience* 12.2, pp. 73–87. ISSN: 1471-003X. DOI: 10.1038/nrn2977.

The ENCODE Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247.

The UniProt consortium (2015). "UniProt: a hub for protein information". In: *Nucleic acids research* 43, pp. D204–12. ISSN: 1362-4962. DOI: 10.1093/nar/gku989.

Thibaud-Nissen, F. et al. (2013). *Eukaryotic Genome Annotation Pipeline*. Berthesda. URL: http://www.ncbi.nlm.nih.gov/books/NBK169439/.

Thompson, A., H. H. Zakon, and M. Kirkpatrick (2016). "Compensatory Drift and the Evolutionary Dynamics of Dosage-Sensitive Duplicate Genes". In: *Genetics* 202.2, pp. 765–774. ISSN: 0016-6731. DOI: 10.1534/genetics.115.178137.

Tian, X., D. S. Kang, and J. L. Benovic (2014). "β-arrestins and G protein-coupled receptor trafficking". In: *Handbook of experimental pharmacology* 219, pp. 173–186. ISSN: 0171-2004. DOI: 10.1007/978-3-642-41199-1_9.

Tomizuka, J., S. Tachibanaki, and S. Kawamura (2015). "Phosphorylation-independent Suppression of Light-Activated Visual Pigment by Arrestin in Carp Rods and Cones". In: *The Journal of biological chemistry*. ISSN: 0021-9258. DOI: 10.1074/jbc.M114.634543.

Train, C.-M. et al. (2017). "Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference". In: *Bioinformatics* 33.14, pp. i75–i82. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx229.

Treangen, T. J. and S. L. Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions". In: *Nature reviews. Genetics* 13.1, pp. 36–46. ISSN: 1471-0064. DOI: 10.1038/nrg3117.

Tsukamoto, H. et al. (2009). "The magnitude of the light-induced conformational change in different rhodopsins correlates with their ability to activate G proteins". In: *The Journal of biological chemistry* 284.31, pp. 20676–20683. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.016212.

Uhlén, M. et al. (2015). "Proteomics. Tissue-based map of the human proteome". In: *Science (New York, N.Y.)* 347.6220, p. 1260419. ISSN: 1095-9203. DOI: 10.1126/science.1260419.

van de Peer, Y., S. Maere, and A. Meyer (2009). "The evolutionary significance of ancient genome duplications". In: *Nature reviews. Genetics* 10.10, pp. 725–732. ISSN: 1471-0064. DOI: 10.1038/nrg2600.

van de Peer, Y., E. Mizrachi, and K. Marchal (2017). "The evolutionary significance of polyploidy". In: *Nature reviews. Genetics* 18.7, pp. 411–424. ISSN: 1471-0064. DOI: 10.1038/nrg.2017.26.

Vilella, A. J. et al. (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates". In: *Genome research* 19.2, pp. 327–335. ISSN: 1088-9051. DOI: 10.1101/gr.073585.107.

Vinckenbosch, N., I. Dupanloup, and H. Kaessmann (2006). "Evolutionary fate of retroposed gene copies in the human genome". In: *Proceedings of the National Academy of Sciences* 103.9, pp. 3220–3225. ISSN: 1091-6490. DOI: 10.1073/pnas.0511307103.

Viphakone, N. et al. (2012). "TREX exposes the RNA-binding domain of Nxf1 to enable mRNA export". In: *Nature communications* 3, p. 1006. ISSN: 2041-1723. DOI: 10.1038/ncomms2005.

Vishnivetskiy, S. A. et al. (2004). "Mapping the arrestin-receptor interface. Structural elements responsible for receptor specificity of arrestin proteins". In: *The Journal of biological chemistry* 279.2, pp. 1262–1268. ISSN: 0021-9258. DOI: 10.1074/jbc.M308834200.

Vishnivetskiy, S. A. et al. (2011). "Few Residues within an Extensive Binding Interface Drive Receptor Interaction and Determine the Specificity of Arrestin Proteins". In: *The Journal of biological chemistry* 286.27, pp. 24288–24299. ISSN: 0021-9258. DOI: 10.1074/jbc.M110.213835.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti (2013). "Detecting natural selection in genomic data". In: *Annual review of genetics* 47, pp. 97–120. ISSN: 0066-4197. DOI: 10.1146/annurev-genet-111212-133526.

Wachter, A. (2014). "Gene regulation by structured mRNA elements". In: *Trends in genetics : TIG* 30.5, pp. 172–181. ISSN: 0168-9525. DOI: 10.1016/j.tig.2014.03.001.

Wain, H. M. et al. (2002). "Guidelines for human gene nomenclature". In: *Genomics* 79.4, pp. 464–470. ISSN: 0888-7543. DOI: 10.1006/geno.2002.6748.

Wainberg, M., B. Alipanahi, and B. J. Frey (2016). "Does conservation account for splicing patterns?" In: *BMC genomics* 17.1, p. 787. ISSN: 1471-2164. DOI: 10.1186/s12864-016-3121-4.

Wang, Z. et al. (2006). "General and specific functions of exonic splicing silencers in splicing control". In: *Molecular cell* 23.1, pp. 61–70. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2006.05.018.

Washietl, S. et al. (2011). "RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data". In: *RNA (New York, N.Y.)* 17.4, pp. 578–594. ISSN: 1469-9001. DOI: 10.1261/rna.2536111.

Waterhouse, A. M. et al. (2009). "Jalview Version 2–a multiple sequence alignment editor and analysis workbench". In: *Bioinformatics* 25.9, pp. 1189–1191. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp033.

Waterhouse, R. M. et al. (2013). "OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs". In: *Nucleic acids research* 41, pp. D358–65. ISSN: 1362-4962. DOI: 10.1093/nar/gks1116.

Waterston, R. H. et al. (2002). "Initial sequencing and comparative analysis of the mouse genome". In: *Nature* 420.6915, pp. 520–562. ISSN: 1476-4687. DOI: 10.1038/nature01262.

Weber, C., T. B. Schreiber, and H. Daub (2012). "Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells". In: *Journal of proteomics* 75.4, pp. 1343–1356. ISSN: 1876-7737. DOI: 10.1016/j.jprot.2011.11.004.

Weintz, G. et al. (2010). "The phosphoproteome of toll-like receptor-activated macrophages". In: *Molecular systems biology* 6, p. 371. ISSN: 1744-4292. DOI: 10.1038/msb.2010.29.

Whelan, S. and N. Goldman (2001). "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach". In: *Molecular biology and evolution* 18.5, pp. 691–699. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a003851.

Wilke, C. O. and D. A. Drummond (2006). "Population genetics of translational robustness". In: *Genetics* 173.1, pp. 473–481. ISSN: 0016-6731. DOI: 10.1534/genetics.105.051300.

Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *The Annals of Mathematical Statistics* 9.1, pp. 60–62. ISSN: 2168-8990. DOI: 10.1214/aoms/1177732360.

Worth, C. L., S. Gong, and T. L. Blundell (2009). "Structural and functional constraints in the evolution of protein families". In: *Nature reviews. Molecular cell biology* 10.10, pp. 709–720. ISSN: 1471-0072. DOI: 10.1038/nrm2762.

Wu, C.-H., M. A. Suchard, and A. J. Drummond (2013). "Bayesian Selection of Nucleotide Substitution Models and Their Site Assignments". In: *Molecular biology and evolution* 30.3, pp. 669–688. ISSN: 0737-4038. DOI: 10.1093/molbev/mss258.

Wu, T. D. and C. K. Watanabe (2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences". In: *Bioinformatics* 21.9, pp. 1859–1875. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti310.

Wyatt, D. et al. (2011). "Small ubiquitin-like modifier modification of arrestin-3 regulates receptor trafficking". In: *The Journal of biological chemistry* 286.5, pp. 3884–3893. ISSN: 0021-9258. DOI: 10.1074/jbc.M110.152116.

Xiao, K. et al. (2007). "Functional specialization of beta-arrestin interactions revealed by proteomic analysis". In: *Proceedings of the National Academy of Sciences* 104.29, pp. 12011–12016. ISSN: 1091-6490. DOI: 10.1073/pnas.0704849104.

Xiao, N. et al. (2015). "SUMOylation attenuates human β-arrestin 2 inhibition of IL-1R/TRAF6 signaling". In: *The Journal of biological chemistry* 290.4, pp. 1927–1935. ISSN: 0021-9258. DOI: 10.1074/jbc.M114.608703.

Xie, W. et al. (2011). "Improving marginal likelihood estimation for Bayesian phylogenetic model selection". In: *Systematic Biology* 60.2, pp. 150–160. ISSN: 1076-836X. DOI: 10.1093/sysbio/syq085.

Yamaki, K. et al. (1990). "Structural organization of the human S-antigen gene. cDNA, amino acid, intron, exon, promoter, in vitro transcription, retina, and pineal gland". In: *The Journal of biological chemistry* 265.34, pp. 20757–20762. ISSN: 0021-9258.

Yan, B. et al. (2011). "Prolyl hydroxylase 2: a novel regulator of β2 -adrenoceptor internalization". In: *Journal of cellular and molecular medicine* 15.12, pp. 2712–2722. ISSN: 1582-4934. DOI: 10.1111/j.1582-4934.2011.01268.x.

Yandell, M. D. and D. Ence (2012). "A beginner's guide to eukaryotic genome annotation". In: *Nature reviews. Genetics* 13.5, pp. 329–342. ISSN: 1471-0064. DOI: 10.1038/nrg3174.

Yang, J.-R. et al. (2012). "Protein misinteraction avoidance causes highly expressed proteins to evolve slowly". In: *Proceedings of the National Academy of Sciences* 109.14, E831–40. ISSN: 1091-6490. DOI: 10.1073/pnas.1117408109.

Yang, Z. (2007). "PAML 4: Phylogenetic Analysis by Maximum Likelihood". In: *Molecular biology and evolution* 24.8, pp. 1586–1591. ISSN: 0737-4038. DOI: 10.1093/molbev/msm088.

Yang, Z. and R. Nielsen (2002). "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages". In: *Molecular biology and evolution* 19.6, pp. 908–917. ISSN: 0737-4038.

Yang, Z., W. S. W. Wong, and R. Nielsen (2005). "Bayes empirical bayes inference of amino acid sites under positive selection". In: *Molecular biology and evolution* 22.4, pp. 1107–1118. ISSN: 0737-4038. DOI: 10.1093/molbev/msi097.

Ye, K. et al. (2008). "Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting". In: *Bioinformatics* 24.1, pp. 18–25. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm537.

Yeh, S.-W. et al. (2014a). "Local packing density is the main structural determinant of the rate of protein sequence evolution at site level". In: *BioMed research international* 2014, p. 572409. ISSN: 2314-6141. DOI: 10.1155/2014/572409.

Yeh, S.-W. et al. (2014b). "Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure". In: *Molecular biology and evolution* 31.1, pp. 135–139. ISSN: 0737-4038. DOI: 10.1093/molbev/mst178.

Yokoyama, S. et al. (2005). "Elephants and human color-blind deuteranopes have identical sets of visual pigments". In: *Genetics* 170.1, pp. 335–344. ISSN: 0016-6731. DOI: 10.1534/genetics.104.039511.

Yoshida, K. et al. (2017). "A unique choanoflagellate enzyme rhodopsin exhibits light-dependent cyclic nucleotide phosphodiesterase activity". In: *The Journal of biological chemistry* 292.18, pp. 7531–7541. ISSN: 0021-9258. DOI: 10.1074/jbc.M117.775569.

Yue, R. et al. (2009). "β-Arrestin1 Regulates Zebrafish Hematopoiesis through Binding to YY1 and Relieving Polycomb Group Repression". In: *Cell* 139.3, pp. 535–546. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.08.038.

Zallot, R. et al. (2016). "Functional Annotations of Paralogs: A Blessing and a Curse". In: *Life (Basel, Switzerland)* 6.3. ISSN: 2075-1729. DOI: 10.3390/life6030039.

Zhai, W. et al. (2012). "Looking for Darwin in genomic sequences–validity and success of statistical methods". In: *Molecular biology and evolution* 29.10, pp. 2889–2893. ISSN: 0737-4038. DOI: 10.1093/molbev/mss104.

Zhan, X. et al. (2011a). "Crystal structure of arrestin-3 reveals the basis of the difference in receptor binding between two non-visual subtypes". In: *Journal of molecular biology* 406.3, pp. 467–478. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2010.12.034.

Zhan, X. et al. (2011b). "Non-visual arrestins function as simple scaffolds assembling the MKK4–JNK3a2 signaling complex". In: *Biochemistry* 50.48, pp. 10520–10529. ISSN: 0006-2960. DOI: 10.1021/bi201506g.

Zhan, X. et al. (2016). "Peptide mini-scaffold facilitates JNK3 activation in cells". In: *Scientific Reports* 6, p. 21025. ISSN: 2045-2322. DOI: 10.1038/srep21025.

Zhang, G. et al. (2014). *Avian Phylogenomics Project*. URL: http://avian.genomics.cn/en/jsp/database.shtml.

Zhang, H. et al. (2003). "Identification and Light-Dependent Translocation of a Cone-Specific Antigen, Cone Arrestin, Recognized by Monoclonal Antibody 7G6". In: *Investigative ophthalmology & visual science* 44.7, pp. 2858–2867. ISSN: 0146-0404. DOI: 10.1167/iovs.03-0072.

Zhang, J. et al. (2016). "Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences". In: *Bioinformatics* 32.20, pp. 3058–3064. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw370.

Zhang, J., R. Nielsen, and Z. Yang (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level". In: *Molecular biology and evolution* 22.12, pp. 2472–2479. ISSN: 0737-4038. DOI: 10.1093/molbev/msi237.

Zhang, M. et al. (2010). "Loss of beta-arrestin1 and beta-arrestin2 contributes to pulmonary hypoplasia and neonatal lethality in mice". In: *Developmental biology* 339.2, pp. 407–417. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2009.12.042.

Zhang, M. et al. (2011). "Disruption of β-arrestins blocks glucocorticoid receptor and severely retards lung and liver development in mice". In: *Mechanisms of development* 128.7-10, pp. 368–375. ISSN: 1872-6356. DOI: 10.1016/j.mod.2011.07.003.

Zhang, X. et al. (2005). "MED1/TRAP220 exists predominantly in a TRAP/ Mediator subpopulation enriched in RNA polymerase II and is required for ER-mediated transcription". In: *Molecular cell* 19.1, pp. 89–100. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2005.05.015.

Zhang, X., T. G. Wensel, and C. Yuan (2006). "Tokay gecko photoreceptors achieve rod-like physiology with cone-like proteins". In: *Photochemistry and photobiology* 82.6, pp. 1452–1460. ISSN: 0031-8655. DOI: 10.1562/2006-01-05-RA-767.

Zhang, Z., N. Carriero, and M. B. Gerstein (2004). "Comparative analysis of processed pseudogenes in the mouse and human genomes". In: *Trends in genetics : TIG* 20.2, pp. 62–67. ISSN: 0168-9525. DOI: 10.1016/j.tig.2003.12.005.

Zhao, J. et al. (2015). "A generalized birth and death process for modeling the fates of gene duplication". In: *BMC evolutionary biology* 15. ISSN: 1471-2148. DOI: 10.1186/s12862-015-0539-2.

Zhou, X. E., K. Melcher, and H. E. Xu (2017). "Understanding the GPCR biased signaling through G protein and arrestin complex structures". In: *Current opinion in structural biology* 45, pp. 150–159. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2017.05.004.

Zhou, Z. et al. (2016). "Codon usage is an important determinant of gene expression levels largely through its effects on transcription". In: *Proceedings of the National Academy of Sciences* 113.41, E6117–E6125. ISSN: 1091-6490. DOI: 10.1073/pnas.1606724113.

Zhu, B.-H. et al. (2016). "PEP_scaffolder: using (homologous) proteins to scaffold genomes". In: *Bioinformatics* 32.20, pp. 3193–3195. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw378.

Zhu, X. et al. (2002). "Mouse cone arrestin gene characterization: promoter targets expression to cone photoreceptors". In: *FEBS letters* 524.1-3, pp. 116–122. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(02)03014-4.

Zhuang, T. et al. (2010). "Elucidation of inositol hexaphosphate and heparin interaction sites and conformational changes in arrestin-1 by solution nuclear magnetic resonance". In: *Biochemistry* 49.49, pp. 10473–10485. ISSN: 0006-2960. DOI: 10.1021/bi101596g.

Zuniga, F. I. and C. M. Craft (2010). "Deciphering the structure and function of Als2cr4 in the mouse retina". In: *Investigative ophthalmology & visual science* 51.9, pp. 4407–4415. ISSN: 0146-0404. DOI: 10.1167/iovs.10-5251.

Zurlo, G. et al. (2016). "New Insights into Protein Hydroxylation and Its Important Role in Human Diseases". In: *Biochimica et biophysica acta* 1866.2, pp. 208–220. ISSN: 0006-3002. DOI: 10.1016/j.bbcan.2016.09.004.