Chapter 2

General Methods

General Methods

To address the broad research questions of my thesis – the evolution of vision and the evolution of chemokine signalling – I used various bioinformatic methodologies. While detailed methods are described in each respective chapter, several basic approaches were shared amongst the different projects. Phylogenetic methods were applied in all projects, and one project additionally incorporated some analyses of single-cell sequencing data. In this chapter I will provide a basic overview of the methodologies, which will serve as a common foundation for the next chapters.

Phylogenetic analyses

All aims within this thesis required phylogenetic analysis of gene families essential to the biological processes of interest. The main steps common to Chapters 3, 4, and 5 are outlined here.

**Dataset preparation**

***Obtaining starting queries***

The first elementary step involves determining which gene families to explore with phylogenetic studies and to obtain reliable reference sequences to use as starting queries for the analyses. While literature serves as a foundational reference, leveraging pathway databases can ensure comprehensive coverage of essential components, especially when examining expansive pathways. One such pathway database is KEGG, which also provides lists of known homologs for pathway components (Kanehisa 2019; Kanehisa et al. 2021). I utilized KEGG as an initial source for reference sequences in Chapters 3 (evolution of phototransduction and photoreceptor cells) and 4 (evolution of retinol metabolism). For Chapter 5 (evolution of chemokine signalling), the primary database of reference was Guide to Pharmacology Database (Bachelerie et al. 2020). For all projects, a supplementary source for reference sequences was UniProt (Boutet et al. 2016; Poux et al. 2017; The UniProt Consortium 2023).

***Choice of species and species trees***

The comparative analysis of systems and signalling pathways requires the examination of genomes and predicted proteomes across a diverse spectrum of species. Thus, an essential preliminary step is selecting the species that best fit the research context. A primary consideration is determining the appropriate taxonomic sampling based on the research question. For example, in Chapters 3 and 4, primary focus was on early branching animals and closest relatives of animals, reflecting the onset of vision in the early stages of animal evolution. Yet, given the possibly ancient origin of certain components of the pathways under study, it was crucial to incorporate representatives from all major eukaryotic lineages. For this, my primary references were Adl 2019 for eukaryotic classification and Burki 2020 for phylogenetic relationships (Adl et al. 2019; Burki et al. 2020). In contrast, the chemokine signalling system is known only in vertebrates, with some non-canonical components potentially existing in other bilaterians. As such, in Chapter 5, species sampling was limited to animals, with an emphasis on vertebrates, a balanced representation of other bilaterians and a few non-bilaterians for a comprehensive search. Another vital consideration in species selection is the quality of available genomes/proteomes. The quality of the predicted proteome can significantly impact the outcomes and reliability of subsequent bioinformatic analyses. High-quality genomes, which are characterised by high levels of completeness and accuracy offer a more reliable representation of an organism’s genetic blueprint. Errors, contamination or ambiguities in the sequence can lead to false or missed identifications, impacting downstream analyses (Simion et al. 2018; Waterhouse et al. 2018; Manni et al. 2021; Simakov et al. 2022). A hallmark of high-quality proteomes is their completeness. If a gene family is not identified in a species with a high-quality complete proteome, then it likely reflects true absence and not a technical limitation. In certain scenarios, there might be key species essential to the study, that may have a proteome with low level of completeness. To compensate for this, the solution is to incorporate multiple closely related species, thereby amplifying the chances of detecting the presence of specific gene families within that taxonomic lineage. The tool I used to assess the proteome completeness was BUSCO (Benchmarking Universal Single-Copy Orthologs) (Waterhouse et al. 2018; Manni et al. 2021). BUSCO searches the proteomes for a list of genes that are known to be universally present in single copy (the “BUSCO” genes) within a taxon. It scans the dataset using lineage-specific BUSCO profiles built using hidden Markov models (HMMs), statistical models that can capture the patterns in a set of sequences (REF). The choice of lineage for the search depends on the organisms under study. For example, in Chapters 3 and 4 I employed the BUSCO profiles designed for eukaryotes, whereas in Chapter 5, I utilized those tailored for metazoans. By providing the percentage of complete BUSCOs identified in each proteome searched, it offers a quantitative measure of the completeness of a dataset in terms of expected gene content. It also differentiates between complete BUSCO genes found in single versus multiple copies. As BUSCO genes are expected to be found in single copy, a high percentage of multi-copy complete BUSCOs may be an indicator of assembly issues. It also assesses the percentage of fragmented and missing BUSCOs, thereby providing a full picture of the proteome completeness. Combining this rigorous assessment with taxonomic considerations, it was possible to build tailored species databases for each Chapter, ensuring the robustness of the subsequent analyses.

**Phylogenetic analyses**

***Initial sequence similarity-based data mining***

The collected queries can be used to identify within the species database, homologous sequences to be used for the phylogenetic analyses. This “data mining” step can be first approached through sequence similarity methods. For this, I used BLAST (Basic Local Alignment Search Tool) for amino acid sequences (Altschul et al. 1997; Camacho et al. 2009). This widely used tool works by searching for an initial short match between the query and the database sequence, after which it attempts to add adjacent amino acid to extend the hit. As the alignment grows it is scored based on the exactness of the match, the extension stops if the score drops below a certain fraction of the highest score. BLAST retains this local alignment if its highest score has an expected value (e-value) below a user defined threshold (Lemey et al. 2009). The resulting hits are therefore considered to be more similar to each other than would be expected by chance, suggesting probable homology. This is a very powerful tool to narrow down potential homologs from large protein databases. The choice of e-value cut-off is critical, as if it is too loose (high) unrelated sequences may be collected, while if it is to strict (low) potential homologs might be missed. The e-value is influenced by the query length and database size: shorter queries and larger databases increase the probability of random hits; therefore, the e-value will tend to be higher in these cases. Given these complexities and recognizing that an optimal e-value might differ across gene families, in this thesis I adopted a strategy of initiating with relatively loose BLAST searches followed by additional methodologies to further refine the results.

***Optimisation of final gene family datasets***

While BLAST served as the foundational method in all my chapters and is in general a very common tool, additional refinement of the gene families can be obtained by diverse strategies. In this thesis, the strategies employed can roughly fit into two categories: targeted versus large-scale approaches. In the first instance, *ad hoc* information about each gene family of interest is used to refine the search. This strategy was employed in Chapter 3, where I refined my BLAST results by a combination of two targeted approaches. Initially, I re-ran BLAST against SwissProt (Boutet et al. 2016; Poux et al. 2017), a high-quality curated database of annotated sequences, retaining only those sequences that correctly matched the desired gene family within the top hits. Subsequently, I filtered sequences by identifying known protein domains typical of each protein family. Further details can be found in the Methods section of Chapter 3. This is a highly precise strategy and ensuring high confidence results, however, it is time consuming and requires a thorough knowledge of the gene families. The alternative approach employs sequence clustering tools to discern the relatedness among sequences, which is advantageous for broader albeit less targeted comparisons. This approach helps filter out unrelated sequences that were initially identified by BLAST but appear unrelated with the rest of the cluster. It also aids in distinguishing sub-families within a larger superfamily and clarifying connections amongst families previously classified solely by function rather than by evolutionary relationships. Different methods employ this clustering strategy. In Chapter 5, I utilized CLANS (Frickey and Lupas 2004), a tool that simply clusters sequences based on all-vs-all BLAST scores. Conversely, Chapter 4 employed more sophisticated methods that combine various clustering, phylogenetic and network analyses algorithms to define orthogroups of sequences (further details are available in Chapter 4).

***Annotating Sequences***

A useful additional step is to provide annotations to the sequences collected. This is because not all species proteomes came with annotations. And to navigate the big trees or clusters of sequences it is useful to have as many sequences as possible with a “name”. The most commonly used method throughout this thesis is to blast sequences vs swissprot and keep the best hit. Where necessary, further manual annotations were applied by taking advantage of species-specific or taxon specific databases. Throughout this thesis useful species or taxon specific databases used were: genecards for human, xx for mouse, flybase for flies, echinobase for sea urchins and other echinoderms and yy for Arabidopsis thaliana.

***Multiple sequence alignment and trimming***

Once final gene families have been optimally curated. The next step is to align the sequences. The idea is that if the sequences are homologous they can be aligned in such a way that each column of the alignment represents a homologous position in a consensus sequence. Where all or most sequences align perfectly there is high conservation of the sequence, where alignment is less good there are more divergent portion of the sequence. And finally there can be gaps throughout the alignament that can either represent true signal or in some cases may be due to technical difficulties. In any case, before moving forward with phylogenetic tree construction, it is useful to remove potentially less useful positions in the alignment. Meaning, if there are too many gaps in a specific position of the alignment this could be problemetic in the tree and contemporarily increase computation times without necessarily adding much phylogenetic information. Conversely some gaps can be useful information. Therefore we use tools that can automatically decide cutoff for trimming alignments. And these can be further tailored according to need. e/g/ if we have short sequences to start with then we don’t want to trim too much if not we are not left with enough phylogenetic signal for the trees.

***Gene tree construction***

Once the msa is ready it can be used for tree construction. All trees in this thesis were built under maximum likelihood. Generally, a model finder was run to determine the best fit tree. Supports for the tree were generally UFB, however, TBE trees were used in chapter 5.

Gene tree to species tree reconciliation

In some cases, it can be useful to re-infer a gene tree on the basis of the knowledge of the species relationships. This can be done using generax.

Single-cell sequencing analyses

For one of my aims – understanding the molecular setup of photoreceptor cells (Chapter 3) I used also single cell sequences analyses of publicly available data. This was because, after determining the presence/absence of phototransduction genes in the genome of target species, I wanted to know whether they were co-expressed within the same cell type – a candidate PRC cell. Furthermore, I wanted to identify other genetic patterns common to animal PRCs. By analysing sc data you can see if there are other genes in common in the cell type. With focus on regulatory genes.

**Preliminary steps**

Choice of datasets / considerations... although generic because details should be in the chapter.

-Species choice

-Different life stages of tissues : really a limitation rather than a choice.

-downloading of the datasets: in my case what I needed was the gene to cell matrix. Say quickly what must have been the steps that the authors used previously to get to this step.

**Clustering cells into metacells**

Choice of method to cluster cells:

-First some background knowledge of options

-Then talk about my choice: metacells because of low coverage...

Issue of having to compare distantly related species

-As well as not always same tissue. E.g. in some cases whole body in others the retina.

Tailored methods / question driven methods – so see more details in respective chapter (chapter 3).

References

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology* [Internet] 66:4–119. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jeu.12691

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* [Internet] 25:3389–3402. Available from: https://doi.org/10.1093/nar/25.17.3389

Bachelerie F, Ben-Baruch A, Burkhardt AM, Charo IF, Combadiere C, Förster R, Farber JM, Graham GJ, Hills R, Horuk R, et al. 2020. Chemokine receptors (version 2020.5) in the IUPHAR/BPS Guide to Pharmacology Database. *IUPHAR/BPS Guide to Pharmacology CITE* [Internet] 2020. Available from: http://journals.ed.ac.uk/gtopdb-cite/article/view/5178

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. Methods in Molecular Biology. New York, NY: Springer. p. 23–54. Available from: https://doi.org/10.1007/978-1-4939-3167-5\_2

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The New Tree of Eukaryotes. *Trends in Ecology & Evolution* [Internet] 35:43–55. Available from: https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(19)30257-5

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10:421. Available from: https://doi.org/10.1186/1471-2105-10-421

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* [Internet] 20:3702–3704. Available from: https://doi.org/10.1093/bioinformatics/bth444

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Science* [Internet] 28:1947–1951. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3715

Kanehisa M, Sato Y, Kawashima M. 2021. KEGG mapping tools for uncovering hidden features in biological data. *Protein Science* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4172

Lemey P, Salemi M, Vandamme A-M eds. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd ed. Cambridge: Cambridge University Press Available from: https://www.cambridge.org/core/books/phylogenetic-handbook/A9D63A454E76A5EBCCF1119B3C56D766

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols* [Internet] 1:e323. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.323

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* [Internet] 33:3454–3460. Available from: https://doi.org/10.1093/bioinformatics/btx439

Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O’Connell BL, Dear P, Martinez DE, Steele RE, et al. 2022. Deeply conserved synteny and the evolution of metazoan chromosomes. *Science Advances* [Internet] 8:eabi5884. Available from: https://www.science.org/doi/10.1126/sciadv.abi5884

Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, Philippe H, Telford MJ. 2018. A software tool ‘CroCo’ detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology* [Internet] 16:28. Available from: https://doi.org/10.1186/s12915-018-0486-7

The UniProt Consortium. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* [Internet] 51:D523–D531. Available from: https://doi.org/10.1093/nar/gkac1052

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* 35:543–548.