Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Master Thesis Bioinformatics

Chin-Hua Huang

Datum
31.05.2021

Reviewers

Kay Nieselt
(Bioinformatik)
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

Nadine Ziemert
(Biologie/Medizin)
Medizinische Fakultät
Universität Tübingen

Huang, Chin-Hua:
Comparison of Functional Conservation of Orthologs and Paralogs
Master Thesis Bioinformatics
Eberhard Karls Universität Tübingen
Thesis period: von 01.12.2020 bis 31.05.2021

If one can see every characteristics not as itself, then the state beyond all changes can be attained.

– Translated from Diamond Sūtra

Abstract

The methodology of gene annotations is based on "ortholog conjecture" which proposed that orthologs should have higher functional similarity than paralogs do. In recent years, ortholog conjecture has been doubted by several authors and causes a great debates on how to design the algorithms and wet-lab experiments. In this thesis, the ortholog conjecture is examined with the analyses of functional similarity on gene ontology annotations and expression correlation on RNA-seq datasets. The functional similarity analyses show that paralogs tend to have higher functional similarity than orthologs do; In the correlation analyses of expression profiles, it indicates that the average expression correlation of orthologs are lower when the time of divergence of the two species is earlier. Nevertheless, the average expression correlation of paralogs shows an opposite trend with respect to the time of divergence. From these results, it is indispensable to include not only orthologs but also more paralogs and comparative omics data with extensive taxa to draw more rigorous conclusions.

Zusammenfassung

Die Methodologie der Genannotationen basiert auf Ortholog-Vermutung, welche besagt, dass die Orthologe höhere funktionale Ähnlichkeiten haben als welche die Paraloge haben. In den letzten Jahren war Ortholog-Vermutung von vielen Forschern gezweifelt und löste viele großen Debatte darüber aus, wie man die Algorithmen und Experimente der Nasslabore entwirft. In dieser Abhandlung überprüfen wir den Gültigkeitsbereich der Ortholog-Vermutung mit den Analysen der funktionalen Ähnlichkeit auf die Genontologie und der den Dateien von der RNA-Sequenzierung. Die Analysen der funktionalen Ähnlichkeiten zeigt sich, dass die Paralogen dazu neigen, höhere funktionale Ähnlichkeiten als Orthologen zu haben. In den Korrelationsanalysen der Expressionsprofile weist hin, dass je länger die Zeit der Verweigung der zwei Spezies ist, desto niedriger sind die durchschnittliche Expressionskorrelation von Orthologe geworden. Allerdings zeigt sich die durchschnittliche Expressionskorrelation von Paraloge eine umgekehrte Tendenz hinsichlich der Zeit der Divergierung. Es ist untentbehrlich, dass man mehr Paraloge in den vergleichenden Analysen einbezieht, um die konsequenteren Schlussfolgerungen zu ziehen.

Acknowledgements

Firstly, I am deeply indebted to Prof. Dr. Kay Nieselt supervised my work with great care and gave feedbacks when there were problems during the analyses of data. She also handled my organizational issues during the three years of my study. I would like to express great gratitude to Prof. Dr. Nadine Ziemert who agreed to be the co-supervisor of my thesis. Doctoral student Theresa A. Harbig recommended OrthoVenn2 as reference database for orthology, which solved the problems of ID-Mapping causing the lack of usable RNA-seq data. I would also like to thank Martin Lang and doctoral Student Susanne Zabel dealt with the server issues so that we could continue to program on the cloud. Alexander Müller, Christian Resl, Florian Kraus, Andreas Wolf, Dilek Tuncbilek, Jennifer Müller and Mathias W. Paz are good companions of mine and we help each other in the lectures and in the master thesis. Roland Henkel, a good friend of mine in Berlin and my master who gives me his 70-year old wisdom and was an informatician in Staatsbibliothek zu Berlin. Last but not least, I would like to thank my family members in Taiwan that they provided material and spiritual supports during my study especially in the period of the corona-pandemics.

Contents

List of Figures

List of Tables

List of Abbreviations

BLAST	Basic local alignment search tool
BP	Biological process (in Gene ontology)
BSR	BLAST score ratio
CC	Cellular component (in Gene ontology)
DCS	Duplication consistency score
FPKM	Fragments per kilobase of exon model per million reads
GO	Gene ontology
IA	Information accretion
MCL	Markov clustering
MF	Molecular function (in Gene ontology)
MRCA	Most recent common ancestor
RBH	Reciprocal best BLAST hits
RPKM	Reads per kilobases of exon model per million reads
TIA	Total information accretion
TPM	Transcripts per million

Chapter 1

Introduction

Given an unknown gene of a known species, would its functions be closer to a homologous gene in the same species or that in another species? According to a textbook of molecular biology, the function of this unknown gene should be closer to the homologous gene in another species [LBM+08], which is a common belief called "ortholog conjecture" (see section 1.1). Recently, this concept has been challenged by many scientists and become an issue that needs to be addressed in bioinformatics. The speed of annotations from empirical results cannot catch up with the rate of high-throughput sequencing, in particular since the invention of next-generation sequencing has produced more than 410,000 of genomes until March 2021 [MSB+]. There are also experiments that cannot be performed because some of them violate ethics. Some organisms simply cannot be cultured under typical laboratory conditions. These problems can affect the progress and accuracy of curation. If the unknown gene of a species has a pendant no matter in relative species or in the genome of the same species indicated by the literature, the curator assigns the unknown gene the well-studied functions from the pendant. In fact, this issue has been unsettled for several decades. Many life scientists including bioinformaticians attempted to find a definite conclusion. However, like many biological questions unfortunately, it has a twilight zone. Statistics applied to one experimental or annotation data set can only provide the conclusions under certain conditions how the data are prepared, pre-processed and evaluated.

In this chapter, the terms in orthology and the characteristics of speciation and gene duplication events in the course of evolution will firstly be defined and illustrated in detail. There are different approaches of computations of an orthology shared by two genes. Because most of the articles concentrate on comparing the functional similarity and expression correlation of the two genes, the same approaches are adopted here and will be introduced in sections 1.3 and 1.4.

The number of species pairs that are explored until today obviously cannot

fully answered on which kingdoms of life or which circumstances ortholog conjecture can be applied. Besides the traditional species pair human and mouse and species pair mouse and rat but also plants and worms are included here as part of the analyses. The sources of datasets and the computations of each metrics will be explained in chapter 2. In chapter 3 and 4, relevant combinations of metrics and species pairs are at the end analyzed and plotted in order to draw more rigorous conclusions.

The following sections provide brief overviews of concepts involved in the issues addressed in this thesis.

1.1 Ortholog conjecture

The conservation and divergence of function are the two fates of a gene on the path of evolution. A new species can carry the genes that preserve their old functions as well as that load with new functions which shape the unique characters of this species. If the history of a set of genes can be traced back their origin to a speciation or a gene duplication event, these genes are called **ortho-** or **paralogs**, respectively, and they are ortho- or paralogous to each other [Fit00]. Both orthologs and paralogs belong to the definition of **homologs**. The set of ortho- or paralogs is also called **ortho- or paralog group**. Interestingly, ortho- or paralogous relations are not transitive(Figure 1.1). It is commonly believed that orthologs tend to preserve functions and paralogs evolve new functions. A classical example of this believe is the case of α- and β-tubulin, which are paralogs having different molecular actions on GTP but both are shared between many eukaryotic species to assemble the microtubules [LBM$^+$08] [HH03] [RYW00]. To examine the scope of this conjecture, it is yet more rational to detach orthology from con- or divergence of functions.

Mutation causes genetic variation. Natural environment influences the direction of genetic evolution. The disadvantageous or harmful changes in the genetic information will not be preserved in the course of time. This mechanism acts especially strong on housekeeping genes, which mostly are responsible for essential physiological processes in an organism. Some mutations on housekeeping genes can even lead to premature death or infertility. After the speciation event, the two orthologous genes accumulate changes independently in the course of time and the functions of these two genes might hence diverge. On the other hand, a gene duplication event does not necessary mean the divergence of function. There are 3 possible consequences being proposed for gene duplication events in general [Hah09]: 1. Sub-functionalization, the ancestral function is divided between the 2 copies. Every copy executes a subset of functions of ancestral function; 2. Neo-functionalization, a new

function is developed in one of the 2 copies; 3. Non-functionalization, the accumulation of mutations can cause deleterious effect after the duplication, resulting in gene loss. Matthew H. Hahn distinguishes the fates after the gene duplication events in finer details and stated that e.g. the increment of gene dosage after the duplication enhances the gene function but does not necessarily create new function. The salivary amylase gene (*AMY1*) in human is such an example. It is found that human population that consumes starch-rich diets has more copies of *AMY1*. Thus, from this perspective, paralogs can also conserve functions [Hah09].

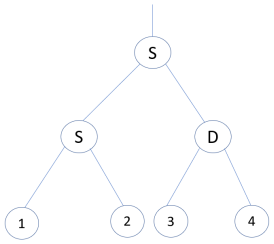

Figure 1.1: The non-transitivity of ortho- and paralogous relationship: The inner nodes marked with S and D are speciation and duplication events, respectively. The leaf nodes marked with numbers are the genes. Both 3 & 1 and 1 & 4 pairs are orthologous, 3 & 4 are paralogous. This demonstrates that orthologous relationships are not transitive. If the node labels of speciation and duplication events are swapped, the same is also true for paralogous relationships.

Historically, the concepts of ortholog and paralog come from a paper of Walter M. Fitch in 1970 [Fit70]. Two genes are "homologous" if both functions are derived from an ancestral gene; Two genes are "analogous" if both functions from two separate genes become similar. Although both definitions cannot be accurately mapped to the modern definitions of ortholog and paralog, Fitch's intention was to prevent the danger of not knowing the relationship between two homologs with similar biochemical functions. He refined the definitions of ortho- and paralog as the one stated above in 2000 [Fit00]. The common belief that gene functions in orthologs are more conserved than those in paralogs is formulated and summarized firstly by Nehrt *et al.* (2011) as "ortholog conjecture"(some of the literature also refers it as **standard model**) and refuted the conjecture based on their evidence in comparative human and mouse microarray data, and gene ontology (GO) annotations that gene functions in paralogs are more similar [NCRH11] than the orthologs do. The opposing views against Nehrt *et al.* will be further introduced in sections 1.3 and 1.4.

1.2 Algorithms and databases predicting orthology

Most of the experiments about evolution and the assignments of gene annotations are conducted according to ortholog conjecture. Prediction of ortho- and paralogs is thus an important task. Tree-based and graph-based methods are 2 main approaches to predict orthology. The former one requires species tree to guide the construction of a gene tree, the methods of which the latter one does not employ. The Ensembl database employs a tree-based method to predict orthology, the pipeline of which can be divided into several steps [VSUV+09]:

1. Consider only the longest protein translation for each gene, then set the edges for the protein pairs with the reciprocal best BLAST hits (RBH) or with BLAST score ratio (BSR) greater than 0.33. BSR can be calculated as follows,

$$BSR = \frac{score(P_1, P_2)}{\max(score(P_1, P_1), score(P_2, P_2))}$$

where $score(P_1, P_2)$ is the BLAST score between protein P_1 and P_2. Every protein pair is assigned with an edge, the weight of which is the BSR. A disconnected graph with many connected components is thus formed.

2. Extract the connected components in the graph from step 1 into clusters. If there is cluster with more than 750 nodes, then repeat step 1 with higher stringency by raising the BSR by 0.1 for each iteration.

3. Perform multiple alignment with MUSCLE for every cluster of proteins.

4. The resulting multiple alignments and a reference species tree are the input of TreeBeST which produces a gene tree. TreeBeST penalizes gene duplications and deletions. This rule is based on the assumptions that gene duplications and deletions are rare and both subtrees of a gene duplication node should be consistent in the course of evolution.

5. Every node v of the gene tree is assigned either as a speciation or gene duplication event. The criterion is set by the duplication consistency score (DCS),

$$DCS(v) = \frac{|S_l(v) \cap S_r(v)|}{|S_l(v) \cup S_r(v)|}$$

where S_l and S_r are the sets of leaf taxa on the left and right branches of node v, respectively. If DCS is less than 0.25 and there is no ortholog, node v will be assigned as a speciation event. The intuition is that both subtrees

should have genes that comes from the gene duplications event within the same taxa, resulting in shared taxa on both subtrees.

6. Calculate the the selection pressure, fraction of non-synonymous substitutions (dN) to synonymous substitutions (dS), for the orthologous gene pairs that diverged within the last 100 million years.

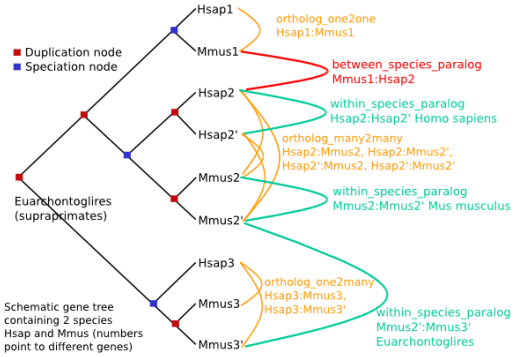

Figure 1.2: Ensembl gene tree [ens]. Gene names(protein names) are on the leaves. Fine grouping of orthology can be inferred by the speciation, gene duplication and whether the two genes are in the same species.

InParanoid is an example of a graph-based clustering approaches [RSS01]. The program accepts 2 proteome FASTA-files. A third proteome file can serve as out-group. The 2 data sets are blasted against itself and against each other, with which 4 asymmetric bit scores. for a protein pair will be obtained. All asymmetric bit scores per protein pair are summarized to an averaged similarity score. The protein pairs with bi-directional best hits (BBH) are set as the centroid of a cluster to which satellite orthologs or in-paralogs from both species are added. This clustering process starts from the protein pairs with high to low sequence identity until a cutoff threshold is reached. If there is an overlap of clusters, they are merged and deleted according to a list of rules.

OrthoMCL develops its pipeline similar to InParanoid but with different clustering approach [LJR03]. It introduces a Markov clustering algorithm (MCL) at the clustering stage. MCL simulates the flows in the BBH-graph. The higher the weight is between an edge, the more possible the flow will be incremented; On the other hand, the edge with less weight will be more probable to vanish, the graph will split ultimately into several connected components. [vD00] Distant paralogous and orthologous relationships can thus be filtered out.

There are web servers integrating the advantages from several orthology prediction methods. One of them is OrthoVenn2 which uses OrthoMCL for orthology clustering [WCDCG15]. The orthology relationships between the clusters are resolved by orthAgogue which uses the multi-threading library of C to improve the efficiency of data parsing, optimize the memory usage

[EKM14]. OrthoVenn2 also develops how to visualize the shared clusters between the compared species. Users can also upload the proteome FASTA of their organisms. Graphical interface can let user choose the organisms to compare. There are also GO-enrichment analyses and interactive network of the resulting clusters attached in the result page. Sequences of the organisms are extracted from the Ensembl database ready for analyses and users to download.

Note that the orthology relationships in Ensembl are finely grouped into many categories (Fig. 1.2) based on the speciation and gene duplication events on the gene tree and whether the two genes are in the same species; For the graph-based algorithm, they can only identify ortho- and paralogs based on the BLAST score and whether the two genes are in the same species. Therefore, there exist different systems of terminology across different articles and algorithms [AGD19]. For simplicity, the two relationships are referred later only as "orthologs" and "paralogs" according to Fitch's definitions.

1.3 Gene Ontology(GO) and functional similarity

All of the gene functions can be summarized in a **ontology**. The term "ontology" can be traced back to the time of Aristotle. "On" means being and "logos" means science or knowledge in ancient Greek, which is the "science of being" [PB08]. In the field of philosophy, ontology asks the following questions: (1) What is an X? (2) Is there any instance(realization) of X? (3) What are the beings that is related to X? [Mei94] One could notice that these questions encompass a recursive structure. If one continues to ask these 3 questions in turns on the things that are related to X, a directed acyclic graph(DAG) can be drawn according to the hierarchy of the relationships which are mostly "is a" or "part of". These 2 predicates point from the concepts with the concrete (**child node**) to the more abstract meaning (**parent node**). In the field of computer science, the ontology means such a graph that organizes the concepts in a certain knowledge field [HRJM15].

Gene ontology contains all the biological processes (BP), cellular components (CC) and molecular functions (MF) in a DAG with BP, CC and MF representing the 3 root nodes. Every node contains a bar code of gene function in the form "GO:" plus 7 digits from 0 to 9. For example, the nodes BP, CC and MF have the bar code `GO:0008150`, `GO:0005575` and `GO:0003674`, respectively. The bar codes are officially called **gene ontology term(GO-term)**. GO-terms can be read and parsed by machines, especially suitable for annotating large amount of genomic data nowadays. Because of the complexity of biochemical interactions, an edge accommodates the

meanings `is_a`, `part_of` and `regulates`, etc. [Dr3].

Because of the acyclic and hierarchical nature of an ontology, we could ask which of the two concepts have the closest semantic meaning. In fact, it is not only just possible but there are hundreds of possibilities to measure the similarity of the two concepts [HRJM15]. The most popular approaches are based on information theory. Philip Resnik proposed a metric in 1995 that calculates firstly the relative frequency of every concept to its root node and then computes the similarity based on the information content of the most recent common ancestor term(MRCA) [Res95]. Dekang Lin proposed the normalization method to improve the Resnik similarity to be independent of the information contents of children under the same MRCA [Lin98]. Gene functions can be seen as concepts. A gene can have many functions and play different roles in different cellular contexts. Therefore, it is necessary to distinguish the **functional similarity** of two genes and **semantic similarity** of two gene functions. Schlicker et al. applied such a concept to compute the functional similarity between the genes and to normalize Lin semantic similarity based on best match average [SDRL06]. Yang-Clark similarity is another functional similarity that views the ontology as a Bayesian network in which the existence of parent node is the prerequisite of the existence of the child node [CR13]. Conditional probability of a child node given its parent node is used to compute the functional similarity instead of the relative frequency of child to root node. (For more details to compute those metrics see chapter 2)

Chen *et al.* (2012) demonstrated that the authorship bias in the GO-based annotations can produce misleading results [CZ12]. High percentage of paralogs in co-study papers in which the same researcher group tend to annotate with the same set of GO-terms. Altenhoff *et al.* (2012). argued similarly that GO annotations could be inferred by different algorithms and each research groups annotates the genes which are specific to their purposes[ASRRD12]. Recent paper by Stamboulian *et al.* (2020) [SGHR20], the authors showed that paralogs have higher functional similarity than orthologs. Stamboulian *et al.* extracted the gene annotations species pair human & mouse and *Saccharomyces cerevisiae & Schizosaccharomyces pombe*, both of which share large evolutionary distances. In this thesis, the approaches of Stamboulian *et al.* are repeated to see if their results still hold true for mouse and rat pair which has smaller evolutionary distance.

1.4 RNA-seq and expression correlations

DNA microarray data used by Nehrt *et al.* has many limitations. There are always irrelevant noises and false positives caused by non-specific hybridization and cross-hybridization. In addition, the exploration of gene expression

is limited only to the available probe sets on the array. RNA-seq is the high-throughput technique using the next-generation sequencing technology and provides the opportunity to detect the expression level across the whole genome. Its pitfall is that there is no standardized protocol for RNA-seq experiments.

The whole process of RNA-seq begins with the extraction of mRNA from the tissue of interest. The extracted mRNA fragments are then fragmented and reverse-transcribed to cDNA which then is ligated with adapter sequences on both ends. There are techniques that focus on obtaining the information if an RNA is sense or anti-sense, which is important for not analyzing the strand-overlapping transcripts. The samples are loaded on next-generation sequencing machine developed by e.g. Illumina. A flow cell with short oligonucleotides complementary to adaptor sequences is prepared for the binding and amplification of cDNAs. dNTP attached with fluorescent label is added for one base at a time during sequencing. The fluorescent labels prevent more than one bases incorporating on to the growing DNA-strand and ensures that one base per strand is called after each release of the fluorescence. Sequencing can be conducted from single end or from both ends. The latter one is beneficial for discovering new transcripts and the analyses of the expression level of different isoforms.

Raw read counts are obtained after the sequencing steps by mapping the RNA-seq reads to the reference genome with the aligner tools which are aware of splice variants (e.g. featureCounts, RTseq and STAR) [dBHS+19]. The aligner tools count how many fragments overlap for each gene. These counts still contain biases that prevent from obtaining the real expression level in the original cells. The longer a gene is, the more reads will be mapped to it. RPKM (reads per kilobases of exon model per million reads), FPKM (fragments per kilobase of exon model per million reads) and TPM (transcripts per million) are three popular metrics that normalize the raw counts.

Definition 1.4.1 *RPKM of a transcript t is defined as*

$$RPKM_t = \frac{10^9 \cdot R_t}{T \cdot L_t}$$

where R_t is the number of reads mapped to the transcript t, T is the number of the total reads and L_t is the length of the transcript t.

FPKM is a variant of RPKM for paired-end reads.

Definition 1.4.2 *TPM of a transcript t can be calculated as follows:*

$$TPM_t = \frac{10^6 \cdot R_t/L_t}{\sum_{i \in I} R_i/L_i} = \frac{10^6 \cdot FPKM_t}{\sum_{i \in I} FPKM_i}$$

where I is the set of all transcripts.

Although gene expression is not equal to function, it is still a proxy to measure the functional similarity between expression profiles across tissues, conditions or several time points [RMSK14]. Chen *et al.* showed that the ortholog conjecture is still supported at least in the case of human and mouse [CZ12]. Most of the works in present days only focus exclusively on the analysis and comparison of the gene expression data of human and mouse. The conclusions of these analyses cannot obviously cover all the species. Thus the RNA-seq data of species pairs built from 4 mammalian species and one avian species (Human, mouse, rat, chimpanzee and chicken), plant species pair (*Arabidopsis thaliana* and *Glycine max*) and round worm species pair (*Caenorhabditis elegans* and *C. briggsae*) are analyzed in this thesis. Here launches an attempt which bridges GO-term analyses with the RNA-seq data set of multiple species and the hope to understand the whole issue on a broader scope.

Chapter 2

Methods and Material

2.1 Sources of data

In this thesis, the gene annotations of species pairs human and mouse, mouse and rat and species pair *Saccharomyces cerevisiae* & *Schizosaccharomyces pombe* have been collected for the analyses of functional similarity. The RNA-seq datasets of species pairs human and chimpanzee, human and mouse, human and chicken, mouse and rat, *Arabidopsis* and soybean and the species pair *Caenorhabditis elegans* and *C. briggsae* have been gathered for the analyses of expression correlations.

2.1.1 Sequence data and files of homology

The web-server OrthoVenn2 computes the files of orthology including orthologs, co-orthologs and in-paralogs. Co-orthologs are combined into orthologs for not complicating the results which mainly focus on examining the validity of ortholog conjecture. The sequences from the two species of interest for computation of orthology are also listed in the result page for download. The source of sequences of OrthoVenn2 comes from Ensembl [WCDCG15].

2.1.2 RNA-seq data

Human, chimpanzee, mouse and chicken RNA-seq datasets have been taken from Brawand *et al.* (2011) [BSN$^+$11]. In that work, they have studied 6 major organs of each organism with both sexes (brain, cerebellum, heart, kidney, liver and testis). In this thesis, the expression values of both sexes are averaged and the expression values of testis are excluded because it is male-specific. Therefore, only 5 organs are considered in this dataset. Comparative mouse and rat RNA-seq datasets come from Söllner *et al.* (2017) and Fushan *et al.* (2015). Both datasets allow to compare the effects caused by the length of

the gene expression profiles. In the work of Söllner *et al* created an expression atlas of mouse and rat with 13 major organs (Brain, Colon, Duodenum, Esophagus, Heart, Ileum, Jejunum, Kidney, Liver, Pancreas, Quadriceps, Stomach and Thymus) [SLH+17]; The work of Fushan *et al.* compares the expression of liver, kidney and brain across 33 mammalian species and attempts to draw conclusions about the relationship between gene expression and life span [FTL+15]. The worm RNA-seq dataset is drawn from Grün *et al.* (2014) who compare the gene expression of *C. elegans* and *C. briggsae* in 13 developmental stages [GKT+14]. Plant RNA-seq dataset originates from Huang *et al.* (2015) who compares the gene expression profiles of 3 root zones(division zone, elongation zone and mature zone) across 7 plants [HS15]. Here, only *Arabidopsis thaliana* and *Glycine max*(soybean) are used for the expression correlation analysis. The unit of the gene expression values of all analyzed datasets is either RPKM or FPKM.

2.1.3 GO-term data

The full gene annotations in UniProt are summarized in a GAF-file which can be obtained from the FTP site of the European Bioinformatics Institute [ebi]. The version in February 2021 is used as analysis of gene functional similarity. In order to map the UniProt accessions to Ensembl accessions used by OrthoVenn2 and the RNA-seq datasets, the mapping files of human, mouse and rat are collected from TSV-repertoire of the Ensembl database and those of the *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are downloaded from TSV-repertoire of the EnsemblFungi database. All mapping files are released in February 2021. The Gene ontology DAG can be extracted from the OBO-file provided by Gene Ontology Resource at the site of citation [obo].

2.2 Implementations for analyses

In this thesis, the quantitative analyses consists of mainly two parts: Expression correlation anayses will be conducted by computing the Pearson correlation coefficient and rank-score similarity of every pair of ortho-/paralogs from the file of orthology computed by OrthoVenn2 for every species pair. Similarly, functional similarity analyses will be directed by calculating the Yang-Clark and Schlicker similarity of every pair of ortho-/paralogs from the file of homology of every species pair. Most of the operations are imported from the package `Pandas`. Most of the concepts are similar to the commands in programming language of relational databases (e.g. SQL). In order to know the interaction between the metrics: sequence identity, expression correlation and functional similarity, each program is written in such a way that it only dealt with two metrics of the three at once. E.g. `semsims.py` computes Functional similarity

and sequence identity; `expr_go.py` is responsible for expression correlation and functional similarity; The programs concerning the expression correlation and sequence identity are written for plants (plants.py), worms (worms.py) and rodents (rodents.py) to adapt different ID formalities in different RNA-seq datasets.

2.2.1 Computation of sequence identity

The procedures of computing the sequence identity are conducted as suggested by Stamboulian et al. (2020) [SGHR20]. All programs mentioned and to be mentioned in this thesis contain the functions of computing the sequence identity which are implemented using the Python Package `Bio`. Pairwise alignments are conducted with BLOSUM62 matrix under the gap opening and extension penalty 11 and 1, respectively. Sequence identity is calculated by dividing the number of character matches in the alignment by the length of the longer protein sequence.

2.2.2 Correlations of expression profiles

The function of Pearson correlation coefficients is imported from the package `Scipy` (`stats.pearsonr`). Pearson correlation coefficients is the one of the most popular metrics used in the expression correlation analysis.

Definition 2.2.1 *The Pearson correlation coefficient of the two gene expression profiles* $\mathbf{x}$ *and* $\mathbf{y}$ *with n conditions is defined as:*

$$r_{\mathbf{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2}}$$

Where $\bar{\mathbf{x}}$ *and* $\bar{\mathbf{y}}$ *are the means of the two gene expression profiles* $\mathbf{x}$ *and* $\mathbf{y}$*.*

From the formula, high correlation coefficients can happen, when both expression profiles have low variance and almost flat parallel to each other on scatter plot. Notice that the Pearson correlation coefficients cannot be defined for 2 complete horizontally parallel lines. For avoiding the high correlations with flat and low expression profiles, the threshold of the variance is chosen to be at least 2 RPKM. Since there are genes with unreasonably high expression values in the datasets, all genes with a mean expression value larger than 9000 RPKM are filtered out. This pre-processing procedure is applied to all the RNA-seq data sets.

Pearson correlation coefficient is sensitive to outliers and nonlinearity . When Chen *et al.* examined whether the ortholog conjecture holds on RNA-seq datasets of human and mouse from Brawand *et al.*, they proposed rank-score similarity of computing the strength of correlation. Rank-score similarity

is a correlation coefficient that is robust against outliers [CZ12]. Firstly, the expression values in each condition are ranked across the genes. The ranked values are converted to percentile ranks. Each ranked expression profile is then normalized with its L^1-Norm which is the sum of the absolute values of all expression values in a ranked expression profile. For convenience, such normalized vector is notated as $\mathbf{r}$.

Definition 2.2.2 *Given two vectors* $\mathbf{r_1}$ *and* $\mathbf{r_2}$ *described above, the rank-score similarity is*

$$rcSim(\mathbf{r_1}, \mathbf{r_2}) = 1 - \|\mathbf{r_1} - \mathbf{r_2}\|_1$$

(The whole computation happens in L^1 *space.)*

2.2.3 Functional similarity of gene annotations

For comparing if different gene annotation metrics can have different effects on the distribution of the functional similarity, Yang-Clark similarity and Schlicker similarity are chosen, both of which are also used in the analyses of Stamboulian et al. (2020) [SGHR20] [CR13]. The gene ontology DAGs for BP, CC and MF are extracted individually from OBO-file with the relationship edges `is_a` and `part_of`. Furthermore, the GO-terms with the evidence codes IC, TAS, IEP, IPI, IGI, IMP, IDA and EXP have been extracted from GAF-file of all gene annotations of UniProt. For each GO-term, paths to its ancestral GO-term are computed by depth-first traversal of the Gene ontology DAG. The paths belong to this GO-term are also called **propagated annotations** [CR13]. Every GO-term in each gene is hashed with its paths. The paths in each gene are then summarized such that there is no repetitively visited node. Note that the OBO-file has already marked the root node to which every GO-term belongs. Therefore, the path hashing process is conducted for BP, CC and MF individually. The first step of computing the Yang-Clark similarity is to determine the conditional probability $P(v|Pa(v))$, i.e. the probability that a GO-term appears given that all of its parent terms appear in the functional annotations of a gene.

Definition 2.2.3 *Information accretion (IA) of a node v is*

$$IA(v) = -log_2 P(v|Pa(v))$$

Because

$$P(v|Pa(v)) = \frac{P(v \cap Pa(v))}{P(Pa(v))} = \frac{n(v \cap Pa(v))}{n(Pa(v))}$$

, the computation of $P(v|Pa(v))$ can be thought as counting firstly the occurrences that the node v and all of its parent term appear together in the annotations of a gene $n(v \cap Pa(v))$, then counting the occurrences that only all of its parent terms appear $n(Pa(v))$. At last, fraction of both counts is taken. This process is conducted by the program `prob_go.py` which produces 3 tables (BP, CC and MF) each containing 2 joint counts described above, the path to root node for every GO-term and the occurrences of every GO-term which will be used in the computation of Schlicker functional similarity (see Definition 2.2.7).

Definition 2.2.4 *Given a set of propagated annotations S of a gene, the total IA (TIA) of the gene is defined as*

$$TIA(S) = \sum_{v \in S} IA(v)$$

Yang-Clark similarity of two sets of propagated annotations S_a and S_b is

$$ycSim(S_a, S_b) = 1 - \frac{(TIA^p(S_a - S_b) + TIA^p(S_b - S_a))^{\frac{1}{p}}}{TIA(S_a \cup S_b)}$$

According to the literature [CR13] [SGHR20], $p = 2$ is chosen, with which the Yang-Clark similarity is equal to the normalized Euclidean distance subtracted from 1.

For the computation of the Schlicker functional similarity, there are some important definitions to be introduced:

Definition 2.2.5 *The Resnik similarity between two GO-terms t_1 and t_2 is defined as*

$$Re(t_1, t_2) = \max_{t \in A(t_1, t_2)} -log_2 P(t)$$

where $A(t_1, t_2)$ is the set of all common ancestors shared between t_1 and t_2, and $P(t) = \frac{n(t)}{n(r)}$, i.e. the relative frequency of GO-term occurrences over the root node occurrences.

Definition 2.2.6 *Lin similarity is the modification of Resnik:*

$$Lin(t_1, t_2) = \frac{-2 \cdot Re(t_1, t_2)}{log_2 P(t_1) + log_2 P(t_2)}$$

Definition 2.2.7 *Schlicker functional similarity between 2 sets of gene annotations S_a and S_b is the best match average of Lin similarities from both annotations:*

$$Sc(S_a, S_b) = (\frac{1}{|S_a| + |S_b|}) \cdot (\sum_{a \in S_a} \max_{b \in S_b} Lin(a, b) + \sum_{b \in S_b} \max_{a \in S_a} Lin(a, b))$$

As it can be seen in the definitions, one firstly must distinguish the similarity between the two GO-terms, then the **functional similarity** between the two genes. The main functions for Yang-Clark and Schlicker functional similarities are implemented using the packages Pandas and Numpy.

Chapter 3

Results

The gene pairs of orthologs and paralogs need to be determined before the expression correlation and functional similarity analyses is conducted. This process is conducted by the metaserver OrthoVenn2 which uses the OrthoMCL to cluster the groups of orthology. The number of orthologs and paralogs predicted by OrthoVenn2 are listed in Table 3.1 as reference. One can see that the species pairs of microorganisms budding and fission yeasts, C. elegans and C. briggsae have approximately the same number of predicted orthologs as that of paralogs; The rest of the eukaryotes have more than several times higher predicted paralogs than orthologs, except for human and chimpanzee. In order to confirm the magnitude of paralogs that human and chimpanzee have, the other combinations of primates are computed. The results show that the primate paralogs are usually scarce e.g. 4131 and 3029 in the species pair Human and Orangutan, Chimpanzee and Orangutan, respectively. The numbers of ortho- and paralogs included in the UniProt gene annotations can be determined by using the files of orthology predicted by OrthoVenn2. The numbers of ortho- and paralogs that can be found for each species pairs in the Uniprot database are listed in Table 3.2. The species pair budding and fission yeasts has more orthologs found in the ontology BP than the number that OrthoVenn2 was predicted, which shows there are redundant gene annotations in this species pair. Although there are more paralogs predicted than orthologs in the genomes of the 3 species pairs by OrthoVenn2, the gene annotations in the UniProt database show the reversed trend: There are generally more orthologs found than paralogs in the gene annotations for every ontology. After the attempts to draw as many species as possible, the species that have enough number of empirical gene annotations for functional similarity analyses are still limited to what Stamboulian *et al.* have used. The Norwegian rat is the only species that Stamboulian *et al.* did not analyse, which can form a species pair with mouse having closer evolutionary relationship.

Before the functional similarity of gene pairs were analysed, the comparison between the metrics of Schlicker similarity and Yang-Clark similarity has been

Species pair	Orthologs	Paralogs
Budding & Fission yeasts	3,284	3,120
C. elegans & *C. briggsae*	16,561	18,094
Arabidopsis & Soybean	49,612	191,196
Mouse & Rat	18,546	34,605
Human & Chicken	12,280	25,377
Human & Mouse	13,772	45,808
Human & Chimpanzee	17,274	4,585

Table 3.1: The numbers of ortho-/paralogs in the proteomes of the species pairs, predicted by metaserver OrthoVenn2.

Species pairs	Ontology	Orthologs	Paralogs
Budding & Fission yeasts	BP	3,332	2,614
	CC	2,344	1,872
	MF	2,500	2,078
Human & Mouse	BP	8,780	5,516
	CC	11,102	8,486
	MF	8,572	5,666
Mouse & Rat	BP	3,436	2,036
	CC	3,312	2,566
	MF	3,014	1,918

Table 3.2: The number of ortho-/paralogs found in the gene annotations in UniProt database for each species pairs.

conducted on the species pair human and mouse (Figure 3.1). In order to make the comparison as fair as possible, the number of orthologs drawn from the gene annotation dataset needs to be the same as the number of the paralogs. In the subplots of the ontology MF for both Schlicker and Yang-Clark similarity, the numbers of drawn samples between orthologs and paralogs are still slightly different after setting the Python program to draw the same number of samples. After several inspections of the programs and the intermediate output files, it is confirmed that there are gene pairs which do not have the GO-term of MF in one of the both genes. This phenomenon can also be observed in the Figure 3.2, which is unique to the ontology MF. For the positions of the distributions of both orthologs and paralogs in the Figure 3.1, the distributions of Schlicker similarity tend to be higher than the Yang-Clark similarity. This is due to the fact that the last step of the computation of Schlicker similarity is based on best average match, which provides less discrimination between the functional similarity of gene pairs. More gene pairs are assigned with the functional similarity 1.0 computed by the Schlicker similarity than by the Yang-Clark similarity. For this reason, the Yang-Clark similarity is applied for rest of the

analyses which are implicated with functional similarity.

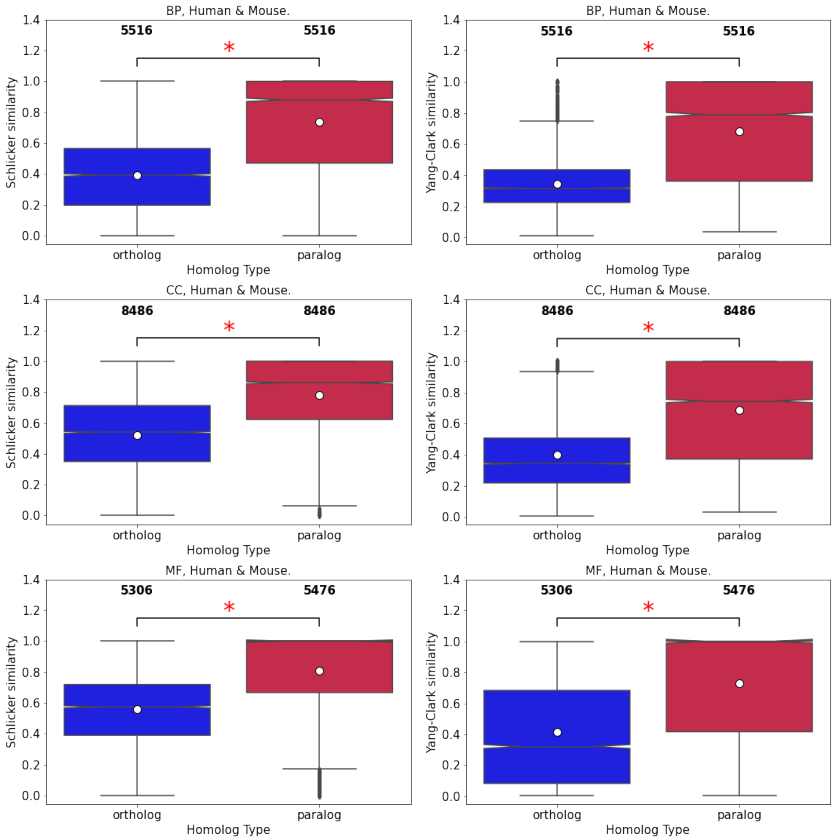

Figure 3.1: The comparison between the distributions that are computed by Schlicker similarity (3 subplots on the right columns) and Yang-Clark similarity (3 subplots on the left columns). Blue and red boxes are the distributions of orthologs and paralogs, respectively. If the notches between the boxes of orthologs and paralogs are not overlapped, the difference is at 95 % of significant level [MTL78]. Every number indicates the number of samples drawn in each bin. Red asterisk with black bar shows the comparison with significant difference according to the result of the independent Student-t test. White dots in each box are the means. The subplots in the figures of this chapter and Appendix follow the same design.

The comparisons of distributions of Yang-Clark similarity in ortho- and paralogs are shown in Figure 3.2. The median and mean of the paralogs in all of the subplots are higher than those of the orthologs. Because the distributions of the ortho- and paralogs are highly skewed in all of the subplots, the median is a better statistics for central tendency. The medians of all species pairs of the orthologs and paralogs in all three ontologies are collected in Table 3.3. It can be seen that the medians of functional similarity of paralogs in MF ontology in all three species pairs are at Yang-Clark similarity 1.0. The Yang-

Clark similarity of all three ontologies of paralogs of mouse and rat are also shown 1.0.

The dependence between the Yang-Clark similarity and sequence identity is examined by plotting the distribution of Yang-Clark similarity for each bin of different level of sequence identity firstly for the ontology BP (Figure 3.3). In most of the bins with comparisons of significant difference in the three subplots, paralogs have higher Yang-Clark similarity than orthologs do, corresponding to the results of Stamboulian *et al.*. The only exceptions of this trend occur at the sequence identity from 0% to 10%, where orthologs sometimes have higher Yang-Clark similarity than paralogs do. In the subplot of ontology BP of species pair budding and fission yeasts, the Yang-Clark similarity of orthologs decreases in the range from 70% to 100% sequence identity. The figures about rest of the two ontologies having similar trends are shown in the Appendix (A.1 and A.2). To further inspect whether there is linear correlation between the metrics Yang-Clark similarity and sequence identity, the Pearson correlation coefficients of the all gene pairs, orthologs and paralogs are computed respectively for each species pair and ontology (Table 3.4). When all gene pairs are considered, the species pair budding and fission yeasts shows the strongest linear correlation among all the species pairs, although it is still moderately correlated. Species pair human and mouse shows almost no linear correlation, and the ontologies BP and CC of mouse and rat show only weak correlation w.r.t. sequence identity. If only the orthologs are considered, there is almost no or at most weak linear correlation. However, if only the paralogs are considered, there are several moderate correlations again among all three ontologies of the species pair budding and fission yeasts and species pair mouse and rat. The linear correlation is higher when only the paralogs are considered than when only the orthologs are considered. In addition, the linear correlation of all gene pairs is mainly contributed by the linear correlation of the paralogs. If the range of sequence identity are limited only to the range from 70% to 100%, the Pearson correlation coefficients of these gene pairs are listed in table 3.5. The only change that can be seen in this table is the more negative Pearson correlation coefficients of orthologs in the species pair budding and fission yeasts in all three ontologies.

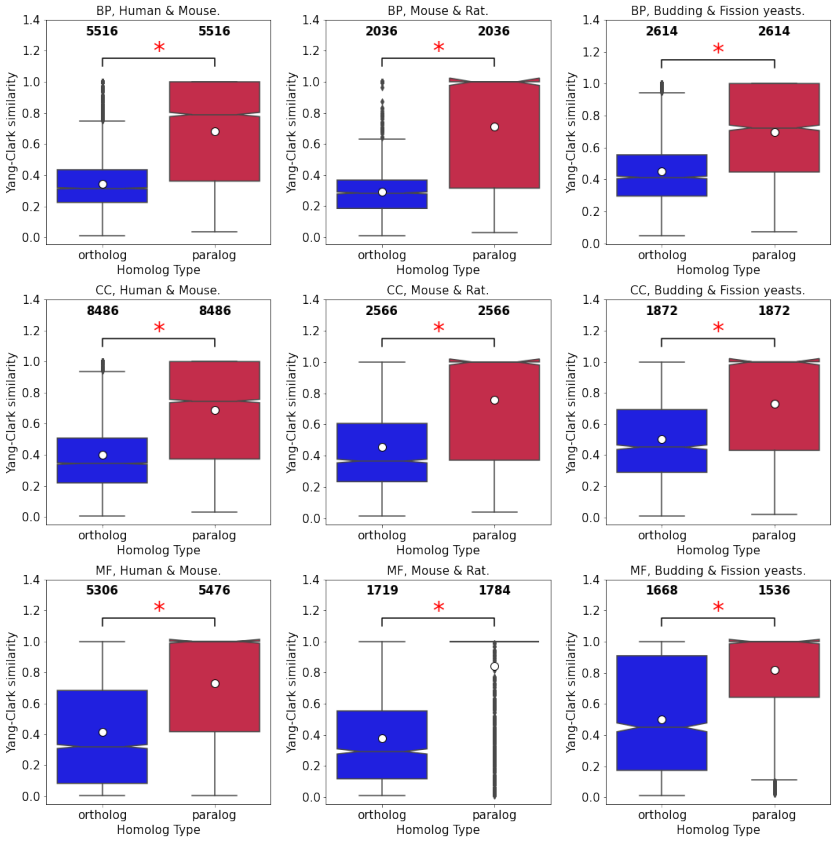

Figure 3.2: The distributions of Yang-Clark similarity of orthologs and paralogs extracted from three different ontologies and the three species pairs extracted from the UniProt database.

Species pairs	Ontology	Orthologs	Paralogs
Budding & Fission yeasts	BP	0.584	0.886
	CC	0.629	0.966
	MF	0.566	1.
Human & Mouse	BP	0.393	0.881
	CC	0.54	0.86
	MF	0.574	1.
Mouse & Rat	BP	0.284	1.
	CC	0.541	1.
	MF	0.442	1.

Table 3.3: The medians of Yang-Clark similarity of orthologs and paralogs of each ontology for each species pairs extracted from UniProt database.

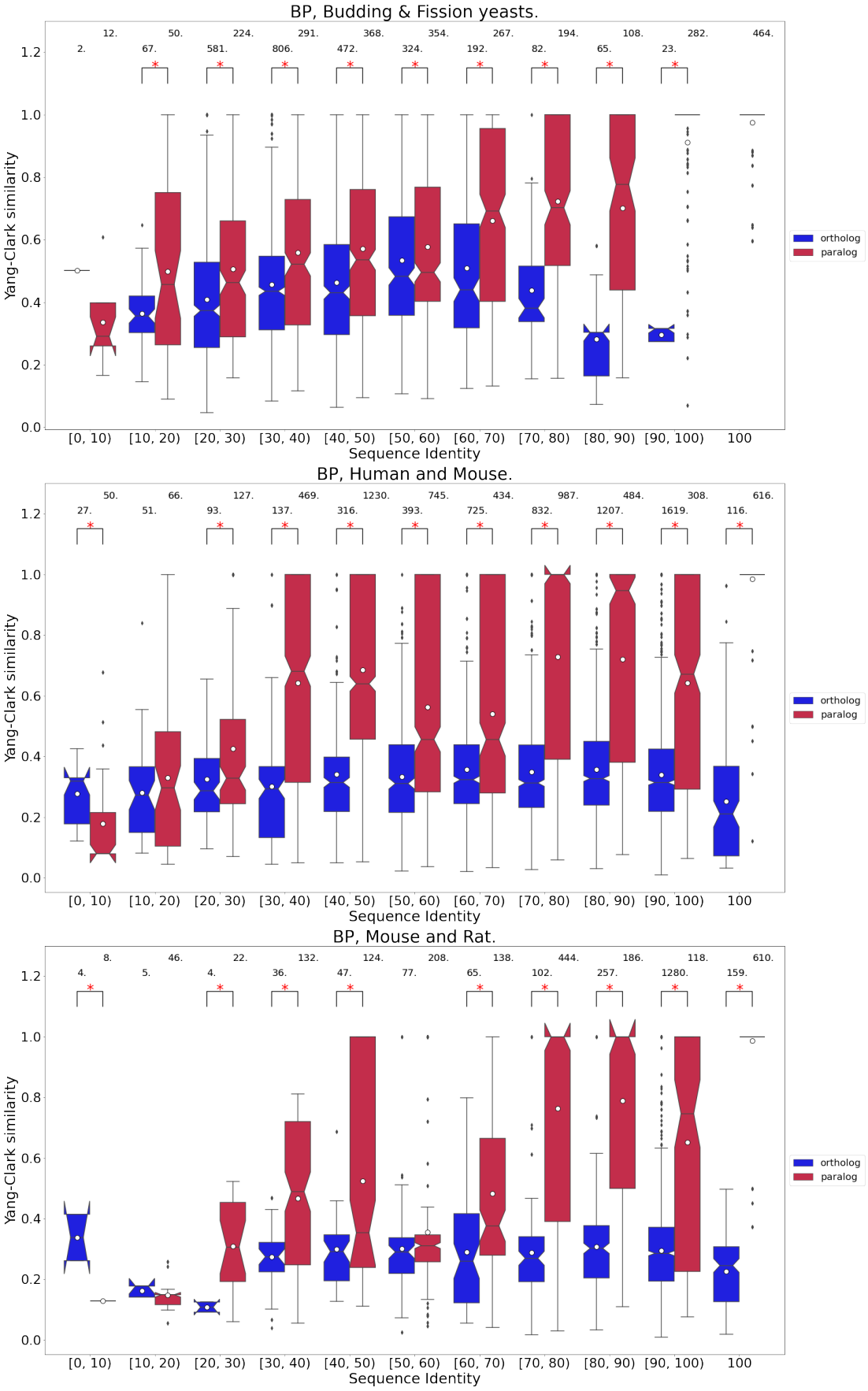

Figure 3.3: The comparison between the box plots of Yang-Clark similarity w.r.t. Sequence identity of three species pairs extracted from the ontology BP.

Species pairs	Ontology	All	Orthologs	Paralogs
Budding & Fission yeasts	BP	0.518	0.054	0.56
	CC	0.408	-0.048	0.467
	MF	0.314	-0.121	0.366
Human & Mouse	BP	0.013	0.022	0.3
	CC	0.033	-0.02	0.351
	MF	-0.026	0.001	0.212
Mouse & Rat	BP	0.148	0.	0.611
	CC	0.234	0.149	0.597
	MF	0.025	-0.097	0.495

Table 3.4: The Pearson correlation coefficient of Yang-Clark similarity w.r.t. sequence identity on all gene pairs, orthologs and paralogs for each species pairs extracted from UniProt database.

Species pairs	Ontology	$SI \geq 0.7$	Orthologs	Paralogs
Budding & Fission yeasts	BP	0.532	-0.382	0.469
	CC	0.56	-0.329	0.519
	MF	0.547	-0.52	0.552
Human & Mouse	BP	0.024	-0.058	0.242
	CC	-0.002	-0.116	0.306
	MF	-0.02	-0.163	0.202
Mouse & Rat	BP	-0.079	-0.032	0.296
	CC	0.002	0.143	0.33
	MF	-0.126	-0.093	0.249

Table 3.5: The Pearson correlation coefficient of Yang-Clark similarity w.r.t. sequence identity on the gene pairs whose sequence identity is greater or equal to 0.7 (the column "$SI >= 0.7$") and on the orthologs and paralogs which have the sequence identity greater or equal to 0.7 are shown for every ontology and species pairs.

On the part of the expression correlation analyses, the numbers of orthologs and paralogs for the comparative RNA-seq datasets used in this work including their authors are listed in the Table 3.6. These numbers are also determined by using the files of orthology predictions of OrthoVenn2. The orthologs included in the comparative RNA-seq datasets are more than the paralogs included in the datasets, with the exception of the species pair *Arabidopsis* and soybean. In contrast to the trend shown in Table 3.1, the numbers of orthologs are several times higher than those of paralogs. Similarly to the functional similarity analyses, before conducting the expression correlation analyses, the effects of different lengths of the gene expression profiles and of the two applied metrics are assessed and illustrated in Figure 3.4. Visually, the two distributions of

orthologs and paralogs computed by the Pearson correlation coefficient are greatly affected by different lengths of the gene expression profiles. Although the effects caused by different lengths of the gene expression profiles are not so obvious in the distributions of rank-score similarity, the positions of the distributions of orthologs and paralogs from Fushan *et al.* (containing 3 organs) are higher than the positions of those from Söllner *et al.* (containing 13 organs). It is because the shorter gene expression profiles are easier to generate high expression correlations and the length of gene expression profiles from the comparative mouse and rat RNA-seq dataset of Söllner *et al.* is much longer than that from the Fushan *et al.*. The reasons that rank-score similarity is more robust than Pearson correlation coefficient are that it is rank-based and considers the length of the gene expression profiles. For this reason, the rank-score similarity is chosen for the following analyses implicated with expression correlations and the comparative RNA-seq dataset of Söllner *et al.* is chosen to represent the results of the species pair mouse and rat.

Species pairs	Authors	Orthologs	Paralogs
C. elegans & C. briggsae	Grün *et al.*	9,289	2,030
Arabidopsis & Soybean	Huang *et al.*	12,318	17,246
Mouse & Rat	Söllner *et al.*	16,600	4,424
Mouse & Rat	Fushan *et al.*	11,186	2,534
Human & Chicken	Brawand *et al.*	8,742	2,328
Human & Mouse		10,524	1,970
Human & Chimpanzee		15,264	1,246

Table 3.6: The numbers of orthologs and paralogs found in the comparative RNA-seq datasets used in this work and their authors.

The numbers of the orthologs and paralogs drawn from the datasets should be equal just as the functional similarity analyses was done, which is shown in Figure 3.5. From the left column of subplots in this figure, it can be seen that orthologs gradually have stronger rank-score similarity than the paralogs do as we move from the top to the bottom. For the right column of the subplots, orthologs have stronger expression correlation than paralogs do. For the available species pairs, it allows to assess the effects of time of divergence on the rank-score distributions of orthologs and paralogs. The median of the rank-score similarity for orthologs and paralogs and the times of divergence are listed in the tables 3.7 and 3.8. In the table 3.7, as the time of divergence decreases, the median of the rank-score similarity of paralogs decreases and that of the orthologs does not show any definite trend. In order to confirm the trend that is influenced by the time of divergence, second group of comparisons is made, which is shown in the table 3.8. In this group of comparison, the human is fixed for every species pair, and the partner of

human is changed from chicken, mouse to chimpanzee which all come from the comparative RNA-seq datasets of Brawand *et al.*. As the time of divergence decreases, the median of rank-score similarity of the paralogs decreases, but that of the orthologs increases.

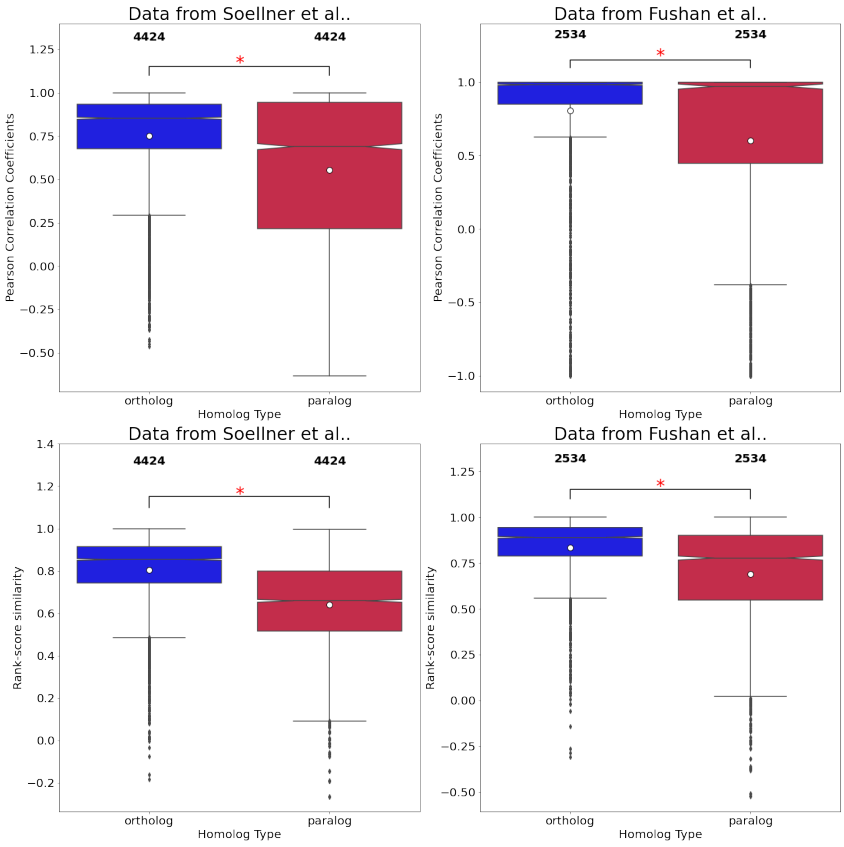

Figure 3.4: The comparison between the distributions of Pearson correlation coefficient (top row of the subplots) and rank-score similarity (bottom row of the subplots) under the effect of long and short gene expression profiles from the data of Söllner *et al.* (left column of subplots) and Fushan *et al.* (right column of subplots), respectively.

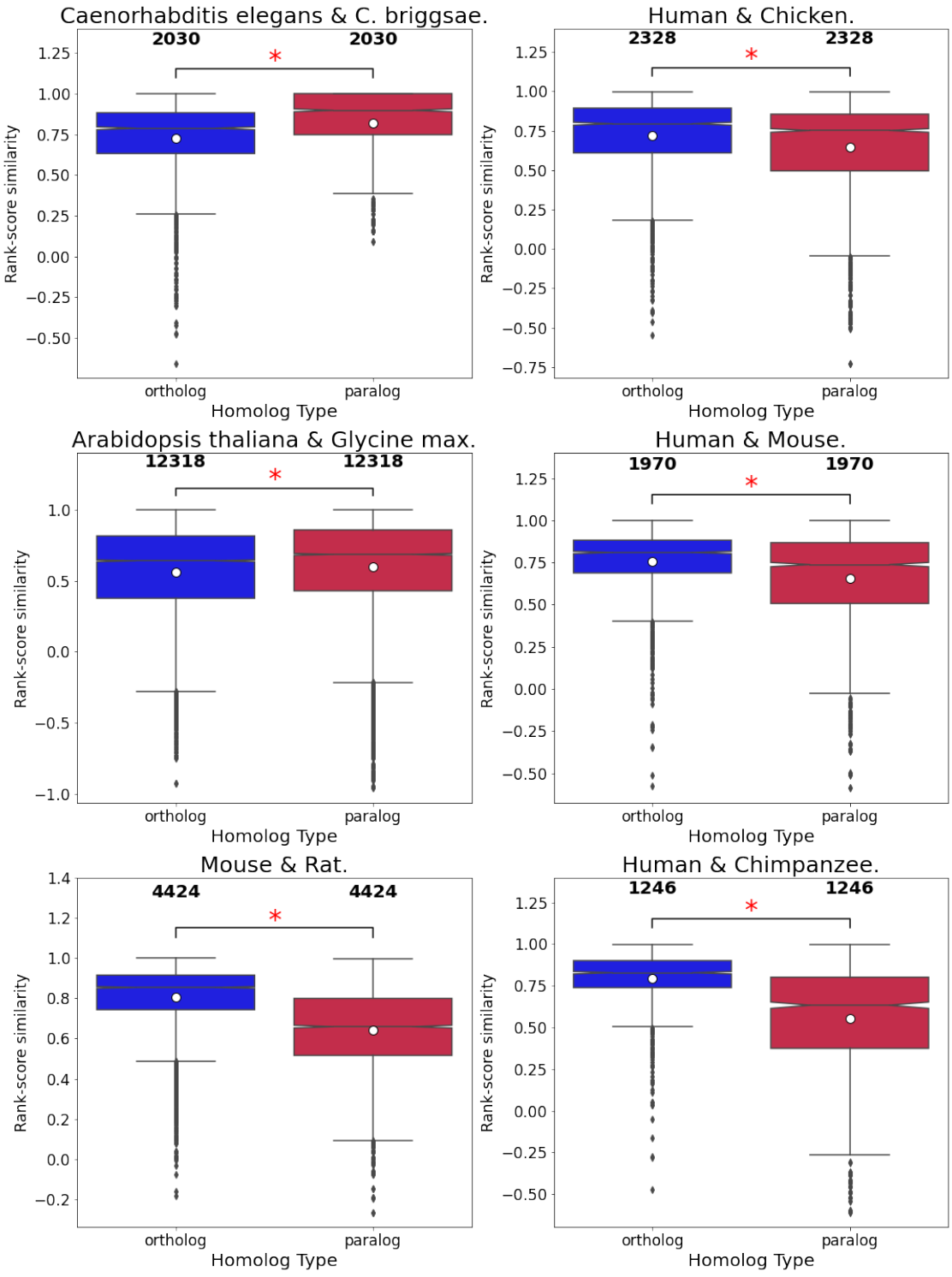

Figure 3.5: The comparison between the distributions of rank-score similarity of all RNA-seq datasets. For each column of subplots, the order of appearance is according to the descending time of divergence. The species pairs for the Table 3.7 and 3.8 are on the left and right column, respectively.

Species pair	Ortholog	Paralog	Time of divergence (Ma)
C. elegans & *C. briggsae*	0.787	0.897	≥ 100
Arabidopsis & Soybean	0.641	0.686	≈ 90
Mouse & Rat	0.854	0.659	≈ 30

Table 3.7: The median of rank-score similarity of orthologs and paralogs for species pairs of *C. elegans* and *C. briggsae*, *Arabidopsis* and soybean and lastly Mouse and Rat, along with their time of divergence (million years ago) [GJC07] [GCS00] [NCL03].

Species pair	Ortholog	Paralog	Time of divergence (Ma)
Human & Chicken	0.795	0.752	≈ 310
Human & Mouse	0.809	0.737	≈ 80
Human & Chimpanzee	0.828	0.633	≈ 6

Table 3.8: The median of rank-score similarity of orthologs and paralogs for each species pair from the RNA-seq datasets of Brawand *et al.*, along with their time of divergence (million years ago) [NCL03] [GN03] [NXG01].

The figures 3.6 and 3.7 show the dependence of rank-score similarity w.r.t. sequence identity as the two groups of comparisons from the two tables 3.7 and 3.8. In the figure 3.6, the paralogs have higher rank-score similarity than the orthologs do in all of the bins of comparisons with significant difference in the species pairs *C. elegans* and *C. briggsae*, *Arabidopsis* and soybean. Orthologs have nevertheless higher rank-score similarity than paralogs do in the species pair mouse and rat. In the group of the species pairs from Brawand *et al.*, the orthologs generally have higher rank-score similarity than the paralogs do. The only exception is at the bin from 10% to 20% sequence identity in the species pair human and chicken where paralogs have higher rank-score similarity than orthologs do. The dependence between the rank-score similarity w.r.t. the sequence identity is summarized into tables A.1 and A.2. According to the two tables, there are only weak correlation among all the species pairs, even if the observed range of the sequence identity of the gene pairs is limited from 70% to 100%.

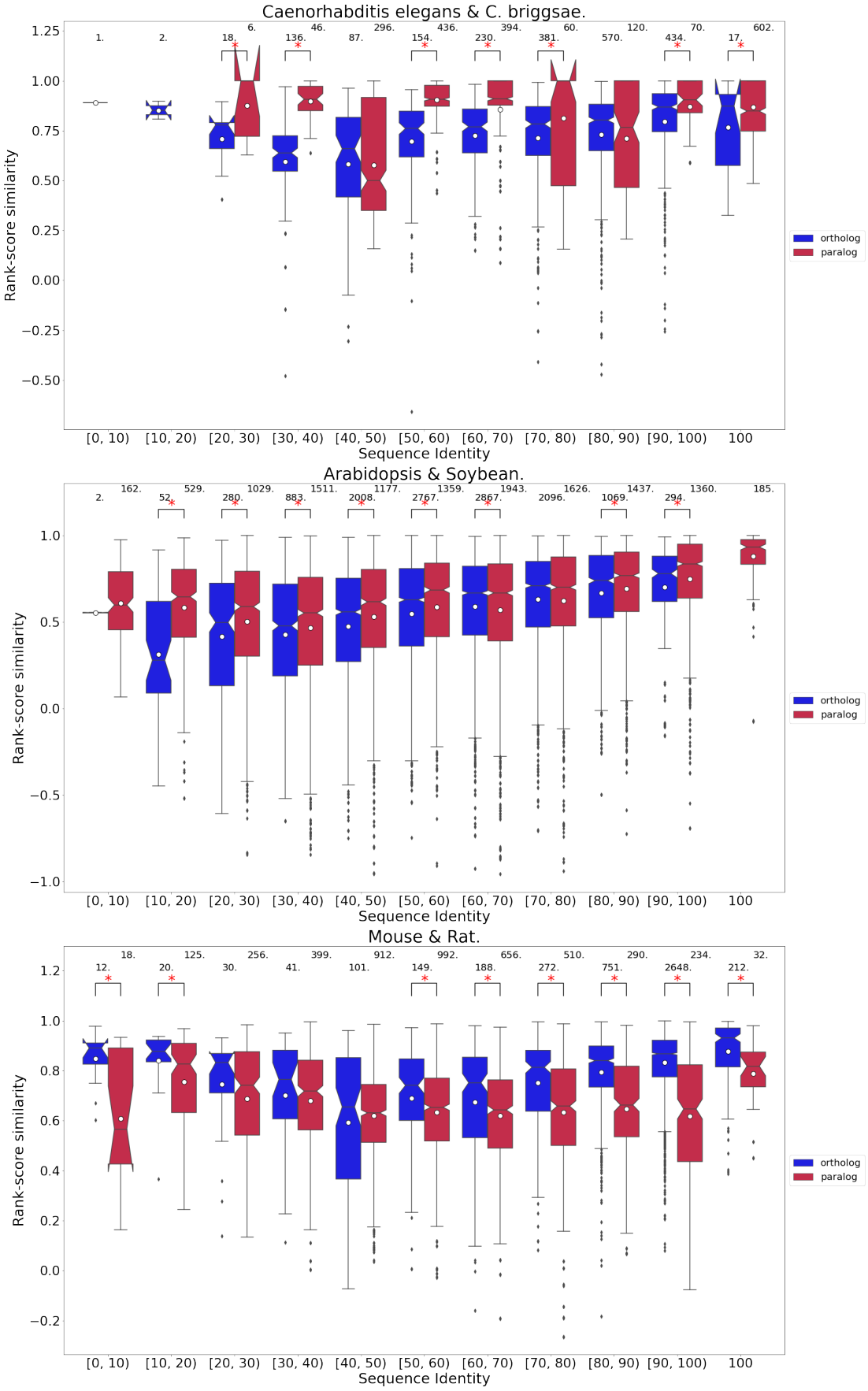

Figure 3.6: The comparison between the distributions of rank-score similarity w.r.t. sequence identity of RNA-seq datasets from Grün *et al.*, Huang *et al.* and Söllner *et al.*. The order of appearance is according to the descending time of divergence.

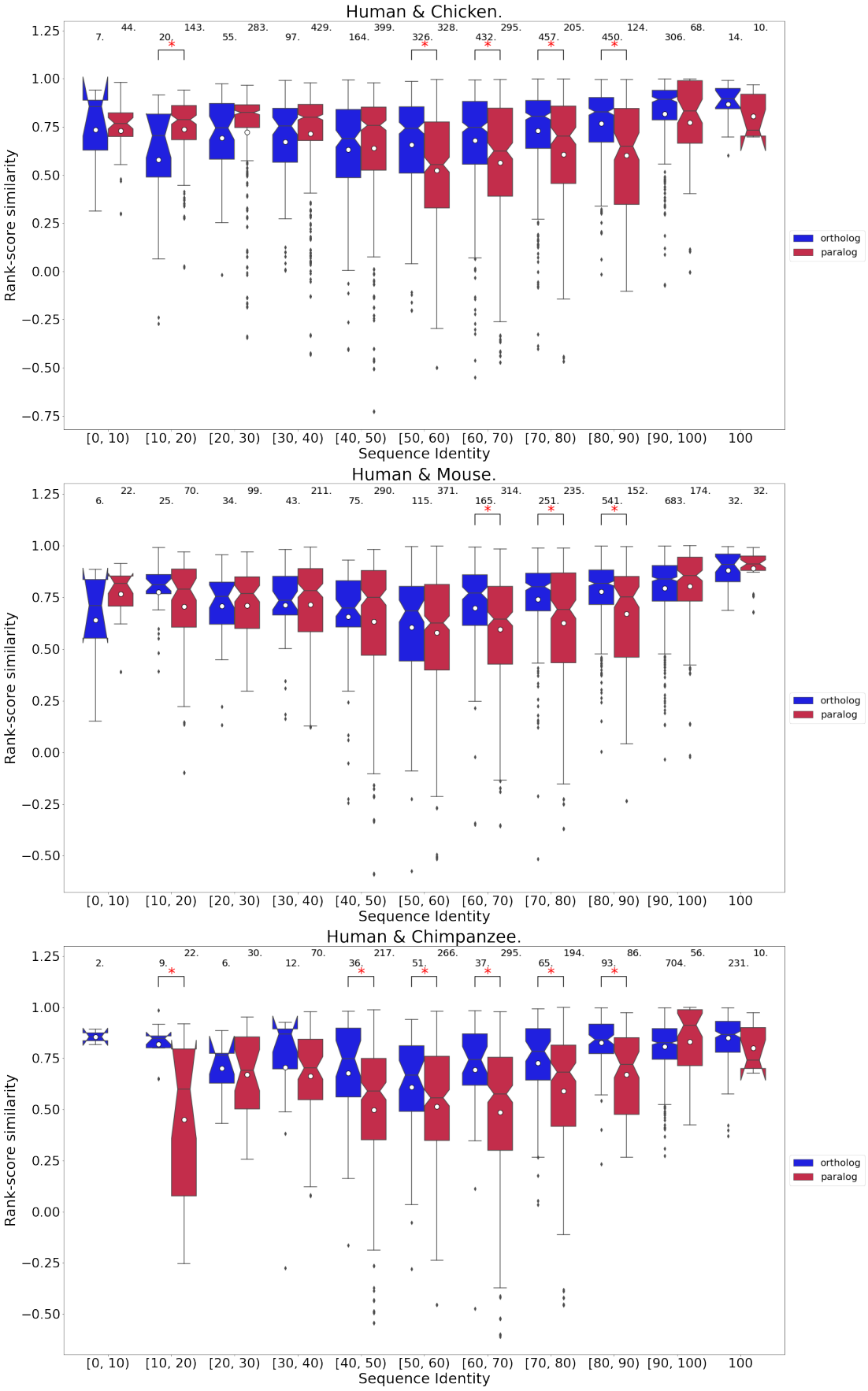

Figure 3.7: The comparison between the distributions of rank-score similarity w.r.t. sequence identity of RNA-seq datasets from Brawand *et al.*. The order of appearance is according to the descending time of divergence.

For the examination of dependence of expression correlation and functional similarity, the numbers shared both in the respective RNA-seq datasets and the UniProt gene annotations are listed in table 3.9. From this table, the numbers of paralogs for each comparative RNA-seq dataset and ontology are much less than the functional similarity analyses, and the numbers of orthologs are several times higher than paralogs. As before, the number of orthologs drawn for the analyses should be the same as the number of the paralogs, but it is expected that there are nuances between the number of orthologs and paralogs drawn for the analyses in the ontology MF. The dependence of the rank-score similarity and Yang-Clark similarity is plotted in the figures 3.8, A.3 and A.4. Using figure 3.8 as an example, there are only weak correlation between Rank-score similarity and Yang-Clark similarity in both orthologs and paralogs. This trend can also be proven by the tables A.3 and A.4. The comparisons with significant difference are concentrated in the range of low Yang-Clark similarity from 0% to 50% for the fact that orthologs are distributed and concentrated at the low functional similarity area.

Species pairs	Ontology	Orthologs	Paralogs
Human & Mouse	BP	4,582	1,050
	CC	6,244	2,860
	MF	4,628	1,218
Mouse & Rat	BP	2,626	550
	CC	2,418	496
	MF	2,152	472

Table 3.9: The numbers of the orthologs and paralogs shared in both the UniProt gene annotations for each ontology and RNA-seq datasets from Brawand *et al.* (human and mouse) or Söllner *et al.* (mouse and rat).

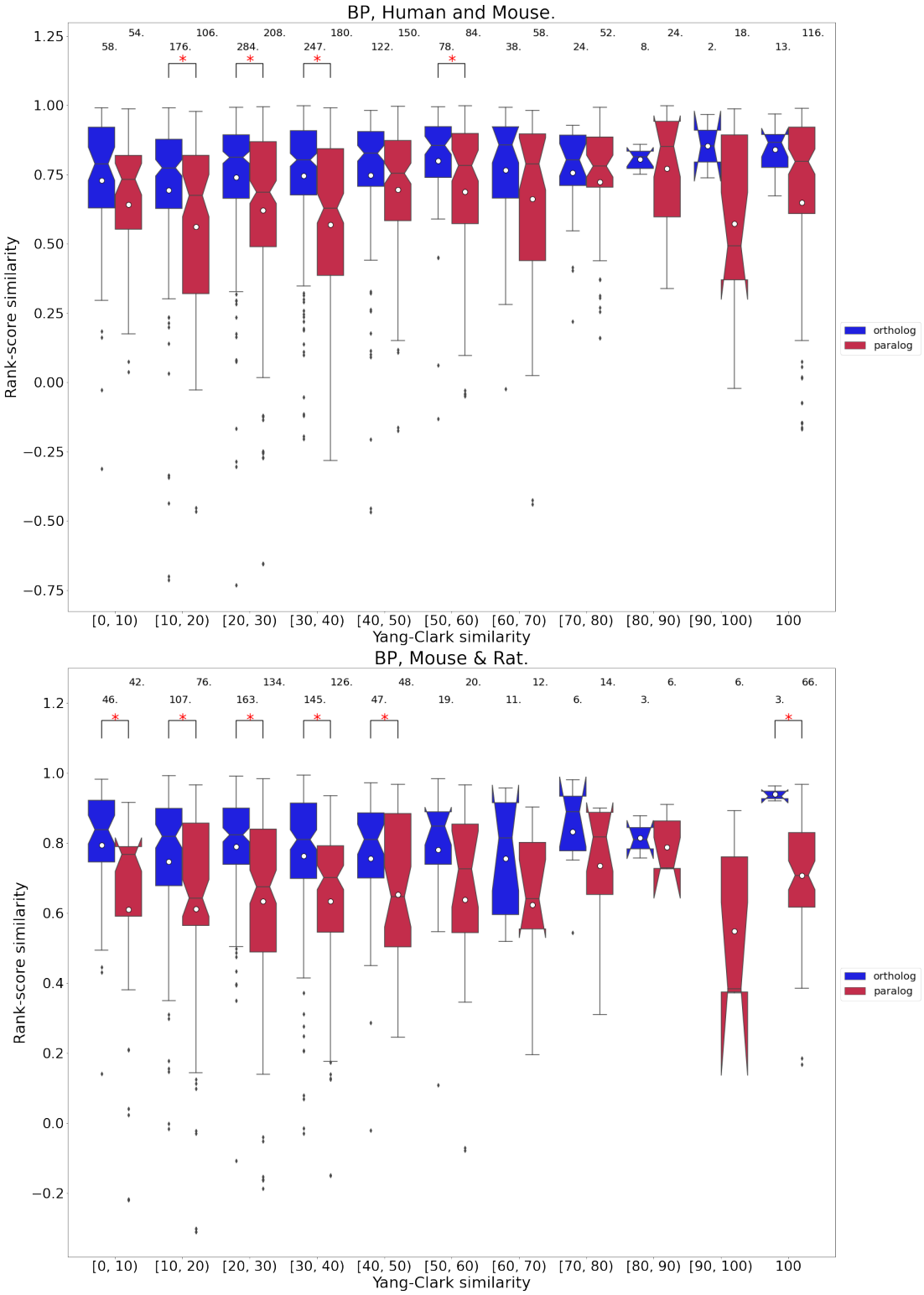

Figure 3.8: The comparison between the box plots of Rank-score similarity w.r.t. Yang-Clark similarity of two species pairs extracted from the gene ontology BP and the respective RNA-seq datasets. Both box plots are arranged according to the descending time of divergence.

Chapter 4

Discussion and Outlook

There have been softwares developed for automatic gene annotations. Their algorithmic designs are based on assigning the gene functions of the orthologs to the unknown genes. Many researchers casted doubt on this approach that whether the assignment of gene function of an orthologous gene are always a optimal strategy to annotate a gene. In 2011, Nehrt *et al.* applied functional similarity analysis on GO-terms and expression correlation analysis on microarray datasets, which established the protocols for the examination of the ortholog conjecture. The functional similarity and expression correlation analyses represent the genetic and phenotypic levels of the similarity measures, respectively. The later works that examined this conjecture could not deviate the two approaches of Nehrt *et al.*, and the approaches can be changed depending on the taxa of the RNA-seq datasets and the databases of the gene annotations examined. After the refutation of the conjecture by Nehrt *et al.*, there were many articles that examined ortholog conjecture. In 2020, Stamboulian *et al.* examined once more the ortholog conjecture with functional similarity analyses and showed once again that the ortholog conjecture does not hold on the gene annotations in the UniProt database. In this thesis, the works of Stamboulian *et al.* (2020) [SGHR20] are examined and extended, including more species pairs and the expression correlation analyses on RNA-seq datasets. Because there is no website for the datasets and implementations of Stamboulian *et al.*. Therefore, a GitHub website has been established for the reference to the RNA-seq and gene annotation datasets and the implementations of different functional similarity and expression correlation metrics used in this thesis [git].

The results of functional similarity analyses are consistent with Stamboulian et al. which show paralogs have higher gene functional similarity than orthologs do. The increment of gene dosage (see Chapter 1) might explain in some cases that duplicated genes may show little or even no change in functions [Hah09]. Moreover, the increment of gene dosage might explain why microorganisms can only have roughly the same number of paralogs as of or-

thologs, not several times higher like eukaryotes. Higher gene dosage can have detrimental effect for microorganisms because of the aggregation of proteins and the loss of their function in such a small volume [RM17].

It is notable to see that paralogs have much higher Pearson correlation coefficient of functional similarity w.r.t. sequence identity than orthologs do. A pair of functionally similar genes that are paralogous tend to have high sequence identity. The reverse of this statement is not true, since similar sequences can also generate different functions at different levels of the central dogma. The fact could be even more complicated if a single cell organism is being discussed because the horizontal gene transfers might be involved. Interestingly, the orthologs of species pair budding and fission yeasts have negative Pearson correlation coefficent of Yang-Clark similarity w.r.t. sequence identity when comparing only those gene pairs that have a sequence identity larger or equal to 70%. As seen in the species pairs human and mouse also mouse and rat, the orthologous genes with the sequence identity of this range should not vary their functional similarity much. The phenomena that are observed in the yeast species pairs could be because either a gene of a third organism are horizontally transferred to these two species pairs and evolves into two different functions in both of them, independently, or a gene is simply transferred between both yeast species. However, these phenomena could be explained by the biological theories only when the gene annotations for these species pairs were complete and empirically curated. The functional similarity analyses in this thesis only include the GO-terms with the evidence codes IC, TAS, IEP, IPI, IGI, IMP, IDA and EXP. According to the statistics provided by the geneontology.org, there are in total $8,006,434$ gene annotations among which $2,008,241$ gene annotations are with the evidence code IEA which means algorithmically annotated without curation and $3,974,367$ are with the evidence code PHYLO which means that they are annotated by the evidence of phylogenetics [sta]. One of the webpages of geneontology.org states that the gene annotations with the evidence code IEA are added mostly by the software InterProScan which infers the function of a gene only on the sequence level, which can cause high sequence identity possibly falsely deducing high gene functional similarity [iea] [FAB17]. In addition, the same webpage states that the annotations with evidence code IEA are assigned based on the one-to-one orthologous relationships in the Ensembl gene trees, which already assumes that orthologs have higher functional similarity than any orthology relationships do.

The results of expression correlation analysis vary between species pairs. The expression correlation of orthologs increases as the time of divergence shortens. On the contrary, the expression correlation of paralogs falls if time of divergence becomes shorter. This indicates that the ortholog conjecture fails when the two species have an early time of divergence. With decreasing time of divergence, the expression correlation in paralogs drops notably. In 2011, Nehrt *et al.* proposed the "cellular context hypothesis" to explain why

paralogs sometimes have higher functional similarity or expression correlation: The evolution of gene functions depends on how the cellular context evolves. Combining the findings of Nehrt *et al.* and the results in this works, the expression correlation might depend on more factors such as the species pairs and their time of divergence. After the speciation event, there might be much longer time for new gene duplication to occur in mouse and chicken than in the chimpanzees, but the same amount of time might also accumulate further changes in the orthologs. The previous works that confirmed the ortholog conjecture analyzed RNA-seq datasets of the mammalian species that mostly show that orthologs have higher expression correlation than paralogs do, which have been confirmed in this thesis by also using the mammalian datasets [CZ12] [RMSK14]. However, the expression correlation of paralogs in the species pair *C. elegans* and *C. briggsae* and the species pair *Arabidopsis* and soybean show that paralogs can have higher expression correlation than orthologs do. The weak mutual dependence of sequence identity, functional similarity and expression correlation show that the magnitude of each metric cannot imply the strength of another.

Most of the species do not possess complete empirical gene annotations in UniProt, and the speed of updates is much slower than the analyses of RNA-seq dataset on the comparisons of multiple species in multiple tissues, conditions and time points. However, the number of comparative RNA-seq datasets still cannot cover all taxonomy of species until today, e.g. fungi, bacteria and other microorganisms, even there are only scarce number of plant comparative RNA-seq datasets available. In this thesis, the work is only concentrated on the eukaryotes partly because nearly all of the bacterial gene annotations are with the evidence code IEA and on account of the difficulty to acquire the comparative bacterial RNA-seq datasets. Another issue of examine the ortholog conjecture with bacterial species pairs is that they have less certain evolutionary history because of the horizontal transfer of genetic materials between the organisms. The definitions of orthology on bacterial species might need to be expanded because the phylogenetic relationships are not based on the tree structure but a network [HB06]. In order to represent horizontal gene transfer, the edges can be drawn between species with distant evolutionary relationships. The orthologs and paralogs are defined by the gene pairs that diverge from the speciation or gene duplication events between the relative species or in the same genome but not the genes that are acquired from a distant lineage of species. These acquired genes from distant species could thus be the confounding factors inside the genomes of a bacterial species pair if only orthologs and paralogs are discussed, especially if the sources of the acquired genes in the examined bacterial species pair are not completely known.

As it was mentioned in Chapter 1 that α- and β-tubulin are paralogs having different molecular mechanisms on GTP but maintaining the microtubules, there is the possibility that orthologous and paralogous gene products interact

with each other and there could be the co-regulations between the two gene expressions. The relationship of functional conservation and orthology could be blurred. This new point of view can change the design of drugs which is still focusing on finding the orthologs of the targeted proteins [CCCS+16]. The search directions can be expanded to the paralogs of this gene and might inhibit the whole pathway of a bacteria, if all of the paralogs form a protein complex of this target. A more extensive benchmark must be developed for the automated annotations on a set of fully curated gene annotations, and it requires not only orthologs but also more paralogs and comparative RNA-seq or proteomics data including a wide range of taxa. The point of view on this issue must be broader, which is not only about whether the ortholog conjecture is valid or not but also why and when the ortholog conjecture has its limitations and how to improve the current approaches to assign gene functions.

Appendix A

Further Tables and Figures

Species pair	All	Ortholog	Paralog
C. elegans & *C. briggsae*	0.183	0.218	0.195
Arabidopsis & Soybean	0.223	0.234	0.225
Mouse & Rat	0.32	0.29	-0.066
Human & Chicken	0.071	0.225	-0.138
Human & Mouse	0.189	0.219	0.042
Human & Chimpanzee	0.396	0.276	0.135

Table A.1: The Pearson correlation coefficients of Rank-score similarity w.r.t. sequence identity for all gene pairs, orthologs and paralogs of each species pairs from RNA-seq datasets used in this work.

Species pair	$SI \geq 0.7$	Ortholog	Paralog
C. elegans & *C. briggsae*	0.272	0.156	0.264
Arabidopsis & Soybean	0.187	0.1	0.214
Mouse & Rat	0.313	0.206	0.023
Human & Chicken	0.214	0.196	0.199
Human & Mouse	0.237	0.132	0.309
Human & Chimpanzee	0.379	0.154	0.287

Table A.2: The Pearson correlation coefficients of rank-score similarity w.r.t. sequence identity on the gene pairs whose sequence identity is greater or equal to 0.7 (the column "$SI >= 0.7$") and on the orthologs and paralogs which have the sequence identity greater or equal to 0.7 are shown for the species pairs from the RNA-seq datasets used in this work.

Species pairs	Ontology	All	Orthologs	Paralogs
Human & Mouse	BP	0.036	0.093	0.087
	CC	0.109	0.061	0.197
	MF	-0.022	-0.002	0.006
Mouse & Rat	BP	0.013	0.018	0.11
	CC	-0.009	0.026	0.078
	MF	0.054	-0.059	0.229

Table A.3: The Pearson correlation coefficients of rank-score similarity w.r.t. Yang-Clark similarity on all gene pairs, orthologs and paralogs for every ontology. The species pairs come from the RNA-seq datasets of Brawand *et al.* (human and mouse) and Söllner *et al.* (mouse and rat).

Species pairs	Ontology	$SI \geq 0.7$	Orthologs	Paralogs
Human & Mouse	BP	-0.133	0.189	-0.133
	CC	0.071	0.1	0.075
	MF	-0.09	0.086	-0.155
Mouse & Rat	BP	-0.101	0.333	-0.041
	CC	0.024	-0.064	0.108
	MF	0.156	-0.099	0.419

Table A.4: The Pearson correlation coefficients of rank-score w.r.t. Yang-Clark similarity on the gene pairs whose sequence identity is greater or equal to 0.7 (the column "$SI \geq 0.7$") and on the orthologs and paralogs which have the sequence identity greater or equal to 0.7 are shown for every ontology. The species pairs come from the RNA-seq datasets of Brawand *et al.* (human and mouse) and Söllner *et al.* (mouse and rat).

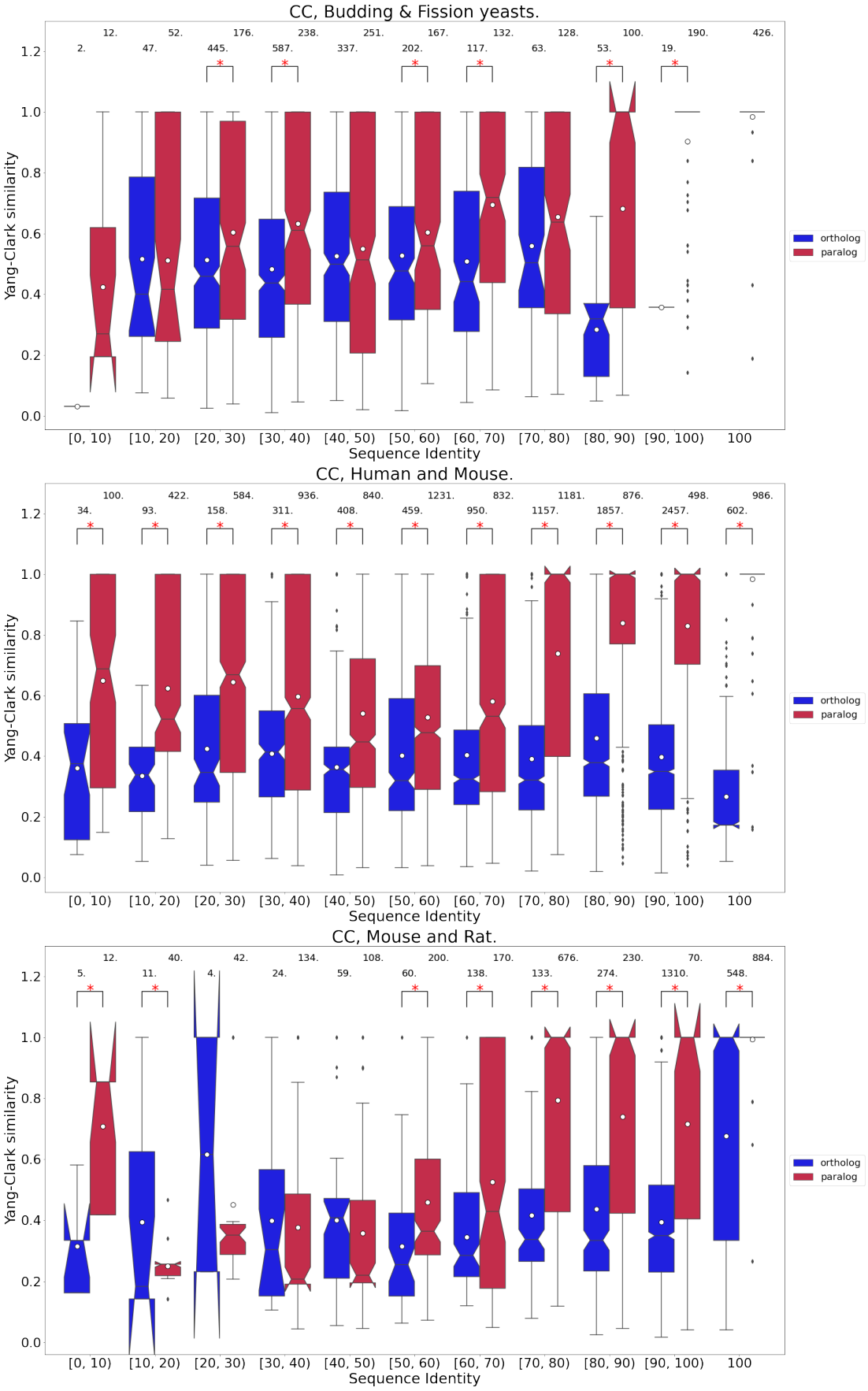

Figure A.1: The comparison between the box plots of Yang-Clark similarity w.r.t. Sequence identity of three species pairs extracted from the gene ontology CC.

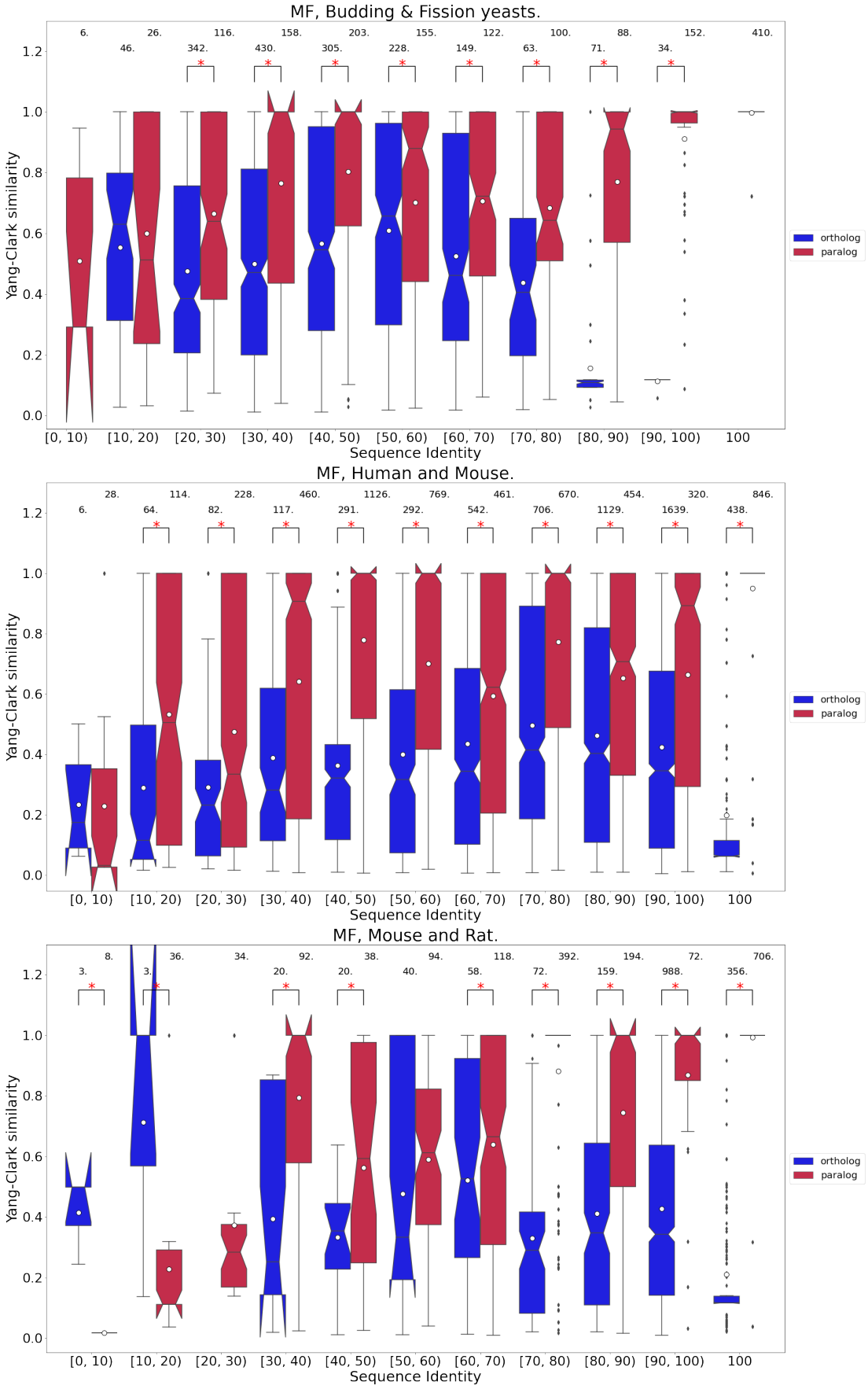

Figure A.2: The comparison between the box plots of Yang-Clark similarity w.r.t. Sequence identity of three species pairs extracted from the gene ontology MF.

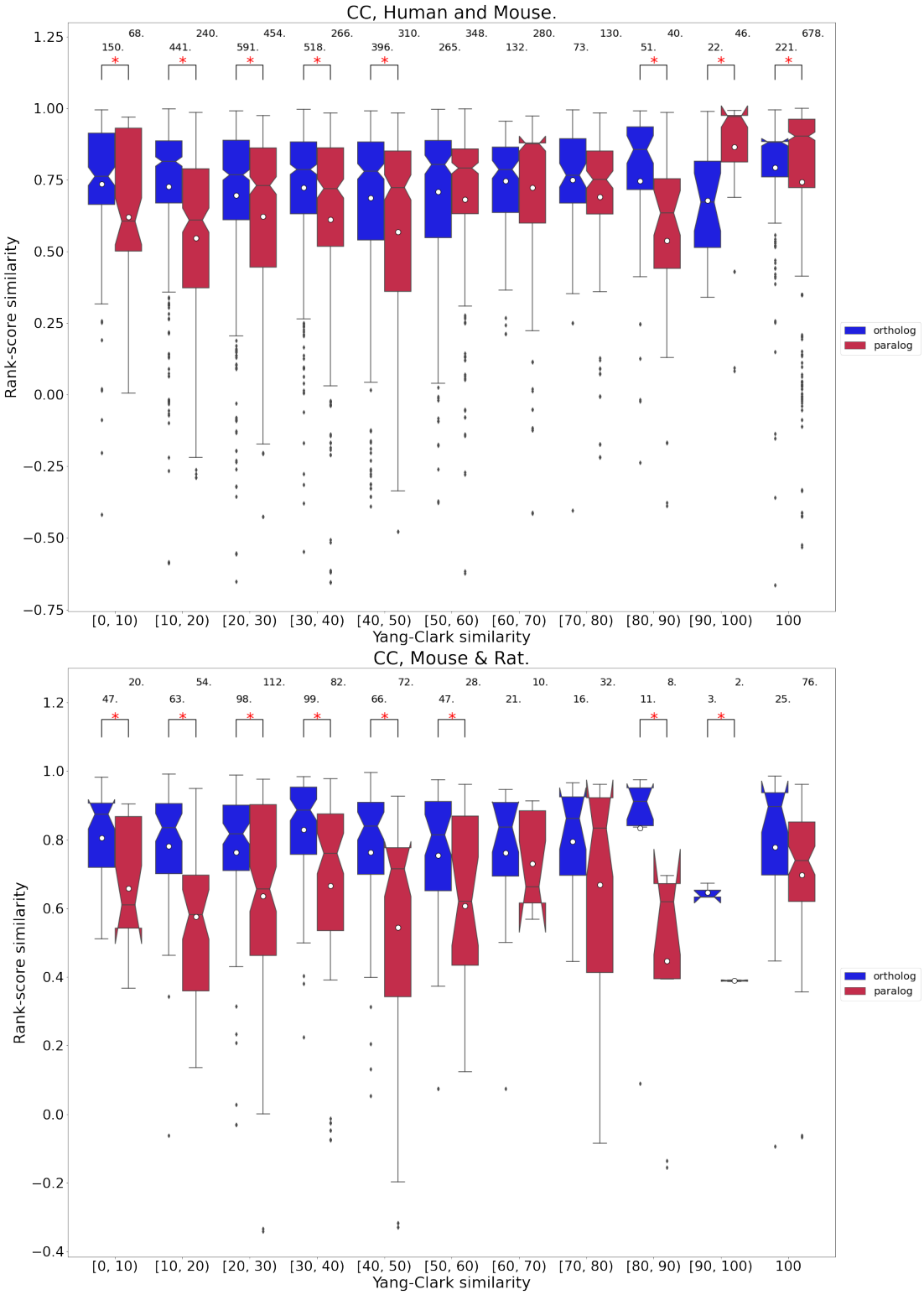

Figure A.3: The comparison between the box plots of Rank-score similarity w.r.t. Yang-Clark similarity of two species pairs extracted from the gene ontology CC and the respective RNA-seq datasets.

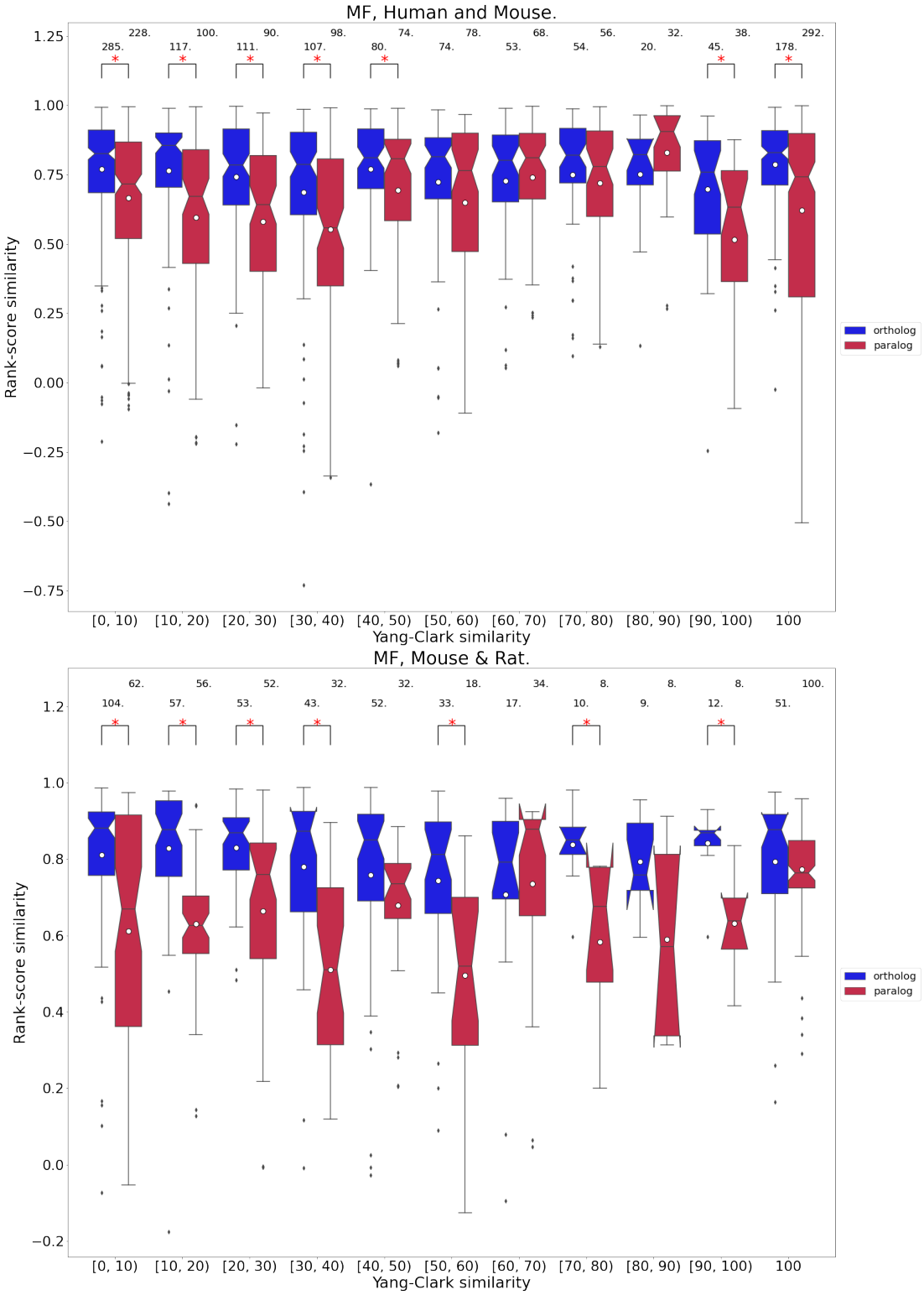

Figure A.4: The comparison between the box plots of Rank-score similarity w.r.t. Yang-Clark similarity of two species pairs extracted from the gene ontology MF and the respective RNA-seq datasets.

Bibliography

[AGD19] Adrian M. Altenhoff, Natasha M. Glover, and Christophe Dessimoz. Inferring orthology and paralogy. *Evolutionary Genomics. Methods in Molecular Biology*, 1910:149–175, 2019.

[ASRRD12] Adrian M. Altenhoff, Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLOS Computational Biology*, 8(5), 2012.

[BSN+11] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, and Svante Pääbo Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478:343–348, 2011.

[CCCS+16] Sriram Chandrasekaran, Melike Cokol-Cakmak, Nil Sahin, Kaan Yilancioglu, Hilal Kazan, James J Collins, and Murat Cokol. Chemogenomics and orthology-based design of antibiotic combination therapies. *Molecular Systems Biology*, 12(872), 2016.

[CR13] Wyatt T. Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29:53–61, 2013.

[CZ12] Xiaoshu Chen and Jianzhi Zhang. The ortholog conjecture is untestable by the current gene ontology but is supported by rna sequencing data. *PLOS Computational Biology*, 8(11), 2012.

[dBHS+19] Koen Van den Berge, Katharina M. Hembach, Charlotte Soneson, Simone Tiberi, Lieven Clement, Michael I. Love, Rob Patro, and Mark D. Robinson. Rna sequencing data: Hitchhiker's guide to expression analysis. *Annu. Rev. Biomed. Data Sci.*, 13(18):139–173, 2019.

[Dr3] Sorin Drăghici. *Data Analysis Tools for DNA Microarray*. Chapman Hall/CRC, 2003.

[ebi] Uniprot goa-file. `http://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz`.

[EKM14] Ole Kristian Ekseth, Martin Kuiper, and Vladimir Mironov. Orthagogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics*, 30(5):734–736, 2014.

[ens] Ensembl: Homology types. `https://www.ensembl.org/info/genome/compara/homology_types.html`.

[FAB17] Robert D Finn, Teresa K Attwood, and Patricia C Babbitt. Interpro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1):D190–D199, 2017.

[Fit70] Walter M. Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19:99–113, 1970.

[Fit00] Walter M. Fitch. Homology, a personal view on some of the problems. *Trends in Genetics*, 16(5):227–231, 2000.

[FTL+15] Alexey A. Fushan, Anton A. Turanov, Sang-Goo Lee, Eun Bae Kim, Alexei V. Lobanov, Sun Hee Yim, Rochelle Buffenstein, Sang-Rae Lee, Kyu-Tae Chang, Hwanseok Rhee, Jong-So Kim, Kap-Seok Yang, and Vadim N. Gladyshev. Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, 14:352–365, 2015.

[GCS00] David Grant, Perry Cregan, and Randy C. Shoemaker. Genome organization in dicots: Genome duplication in arabidopsis and synteny between soybean and arabidopsis. *PNAS*, 97(8):4168–4173, 2000.

[git] Github main site of this thesis. `https://github.com/Huaramo/masterthesis`.

[GJC07] Bhagwati P. Gupta, Robert Johnsen, and Nansheng Chen. Genomics and biology of the nematode caenorhabditis briggsae. *WormBook*, 2007.

[GKT+14] Dominic Grün, Marieluise Kirchner, Nadine Thierfelder, Marlon Stoeckius, Matthias Selbach, and Nikolaus Rajewsky. Conservation of mrna and protein expression during development of c. elegans. *Cell Reports*, 6:565–577, 2014.

[GN03] Galina V. Glazko and Masatoshi Nei. Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution*, 20(3):424–434, 2003.

[Hah09] Matthew W. Hahn. Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5):605–617, 2009.

[HB06] Daniel H. Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–267, 2006.

[HH03] Joe Howard and Anthony A. Hyman. Dynamics and mechanics of the microtubule plus end. *Nature*, 422:753–758, 2003.

[HRJM15] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. *Semantic Similarity from Natural Language and Ontology Analysis*. Morgan Claypool, 2015.

[HS15] Ling Huang and John Schiefelbein. Conserved molecular program for root development in diverse plants. *The Plant Cell*, 27(8):2119– 2132, 2015.

[iea] Information about evidence codes. `http://geneontology.org/docs/guide-go-evidence-codes/`.

[LBM+08] Harvey Lodish, Arnold Berk, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Lawrence Zipursky, and James Darnell. *Molecular Cell Biology*. W. H. Freeman, 2008.

[Lin98] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, 1998.

[LJR03] Li Li, Christian J. Stoeckert Jr., and David S. Roos. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Cold Spring Harbor Laboratory Press*, 13:2178–2189, 2003.

[Mei94] Uwe Meixner. Von der wissenschaft der ontologie. *Logos (neue Folge)*, 1:375–399, 1994.

[MSB+] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy. Genomes online database. `https://gold.jgi.doe.gov/`.

[MTL78] Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[NCL03] Anton Nekrutenko, Wen-Yu Chung, and Wen-Hsiung Li. An evolutionary approach reveals a high proteincoding capacity of the human genome. *TRENDS in Genetics*, 19(6):4168–4173, 2003.

[NCRH11] Nathan L. Nehrt, Wyatt T. Clark, Predrag Radivojac, and Matthew W. Hahn. Testing the ortholog conjecture with comparative functional genomic data from mammals. *Bioinformatics*, 7(6), 2011.

[NXG01] Masatoshi Nei, Ping Xu, and Galina Glazko. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *PNAS*, 98(5):2497–2502, 2001.

[obo] Download ontology. `http://purl.obolibrary.org/obo/go.obo`.

[PB08] Peter Prechtl and Franz-Peter Burkard. *Metzler Lexikon Philosophie*. Verlag J.B. Metzler, 2008.

[Res95] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, 1995.

[RM17] Alan M. Rice and Aoife McLysaght. Dosage-sensitive genes in evolution and disease. *BMC Biology*, 15(78), 2017.

[RMSK14] Igor B. Rogoziny, David Managadzey, Svetlana A. Shabalina, and Eugene V. Koonin. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution*, 6(4):754–762, 2014.

[RSS01] Maido Remm, Christian E. V. Storm, and Erik L. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041–1052, 2001.

[RYW00] Gerald M. Rubin, Mark D. Yandell, and Jennifer R. Wortman. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, 2000.

[SDRL06] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7(302), 2006.

[SGHR20] Moses Stamboulian, Rafael F. Guerrero, Matthew W. Hahn, and Predrag Radivojac. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics*, 36:219–226, 2020.

[SLH+17] Julia F. Söllner, German Leparc, Tobias Hildebrandt, Holger Klein, Leo Thomas, Elia Stupka, and Eric Simon. An rna-seq atlas of gene expression in mouse and rat normal tissues. *Scientific Data(Nature)*, 170185, 2017.

[sta] Statistics of gene annotations. `http://geneontology.org/stats.html`.

[vD00] Stijn Marinus van Dongen. Graph clustering by flow simulation. *PhD thesis, Center for Math and Computer Science (CWI)*, 05 2000.

[VSUV+09] Albert J. Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Journal of Heredity*, 19:327–335, 2009.

[WCDCG15] Yi Wang, Devin Coleman-Derr, Guoping Chen, and Yong Q. Gu. Orthovenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research)*, 43, 2015.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift