Quantitative and qualitative analyses of in-paralogs

Dissertation zur Erlangung des naturwissentschaflichen Doktorgrades der Bayerischen Julius-Maximilians Universität Würzburg

vorgelegt von

Stanislav Vershenya

aus Minsk, Weißrussland

Würzburg 2010

Eingereicht am:

Mitglieder der Promotionskomission:

Vorsitzender: Prof. Dr. M.J. Müller

Gutachter: Prof. Dr. J. Schultz

Gutachter: Prof. Dr. J. Tautz

Tag des Promotionskolloquiums:

Doktorurkunde ausgehändigt am:

Erklärung

Hiermit erkläre ich ehrenwörtlich, daß ich die vorliegende Dissertation selbsständig angefertigt und keine anderen als die angegeben Quellen und Hilfsmittel verwendet habe.

Die Disseertation wurde bische in gleicher noch änlicher Form in einem anderen Prüfungsverfahren vorgelegt.

Außer dem Diplom in Medizin von Weißrussische Medizinische Staatsuniversität und Dr. med. von der Bayerischen Julius-Maximilians Universität Würzburg habe ich bisher keine weiteren akademischen Graden erworben oder versucht zu erwerben.

Würzburg, April 2010 Stanislav Vershenya

I dedicate this work to my family and thank them for their support and love trough all my studying years

Contents

List of Figures

List of Tables

1 Introduction

1.1 Comparative analysis of species

In the pre-genome era, comparison of species relied mainly on anatomical and behavioural differences. With the upcoming of fully sequenced genomes, I now can extend this comparison to the genome sequence. Whereas the first analyses mainly aimed for the detection of commonalities between widely divergent species, with the sequencing of more and closely related species the focus nowadays is more on the detection of differences between the genomes. The hope is to understand the mechanisms underlying morphological, physiological, ecological differences.

A comparison of the mouse and human genome revealed, for example, gene clusters in mouse, indicating species specific gene duplication. Functional analysis revealed that most of these involved genes are involved in reproduction and immunology (Waterston *et al.*, 2002).

1.2 Fate of duplicated genes

In my research I am looking more generally on gene duplications to understand the origin and the long term faith of duplicated genes observed in today's species.

Gene duplications range from single gene duplications to the duplication of the entire genomes (Kellis *et al.*, 2004). Accordingly, about 30-60% of the genes in eukaryotic genomes arose via duplication (Ball and Cherry, 2001). Still, one can expect that the vast majority of new duplicates are destined for extinction. Most never reach appreciable population frequencies and, of those that do, most suffer degenerative mutations that render them non-functional pseudogenes.

The survival of a duplicated gene hinges on whether it provides an evolutionary advantage to the organisms, for example by evolving a new function through fixing the beneficial mutations (neofunctionalization) before being silenced by degenerative ones. However, this part of evolution of duplicated genes is not the most common one.

What happens more often seems to be relaxation of purifying selection immediately after duplication, resulting in accelerated evolution in both duplicated genes. Two new gene copies become fixed for degenerative mutations at complementary subfunctions such that both gene copies are required to cover the multiple subfunctions once performed by the parent gene (subfunctionalization) (fig. 1.1) (Force *et al.*, 1999; Lynch and Conery, 2000; Lynch and Force, 2000; Kondrashov *et al.*, 2002).

It is tempting to argue that neofunctionalization occurs less often, as it depends on a rare class of beneficial mutations, whereas subfunctionalization depends on an abundant class of degenerative ones. But the probability of subfunctionalization may not be as high as it at first seems, as it strongly depends on the number of independently mutable subfunctions that new gene duplicates have in common (Force *et al.*, 1999), and several whole-genome surveys have revealed that the number of common subfunctions is often limited from the start.

There is also a possibility of combination of subfunctionalization and neofunctionalization. It was discovered by examining protein-protein interactions of paralogous gene products in yeast (He and Zhang, 2005). Model suggests a more complex subneofunctionalization model under which the evolution of paralogs starts with rapid subfunctionalization but subsequently often switches to the neofunctionalization mode.

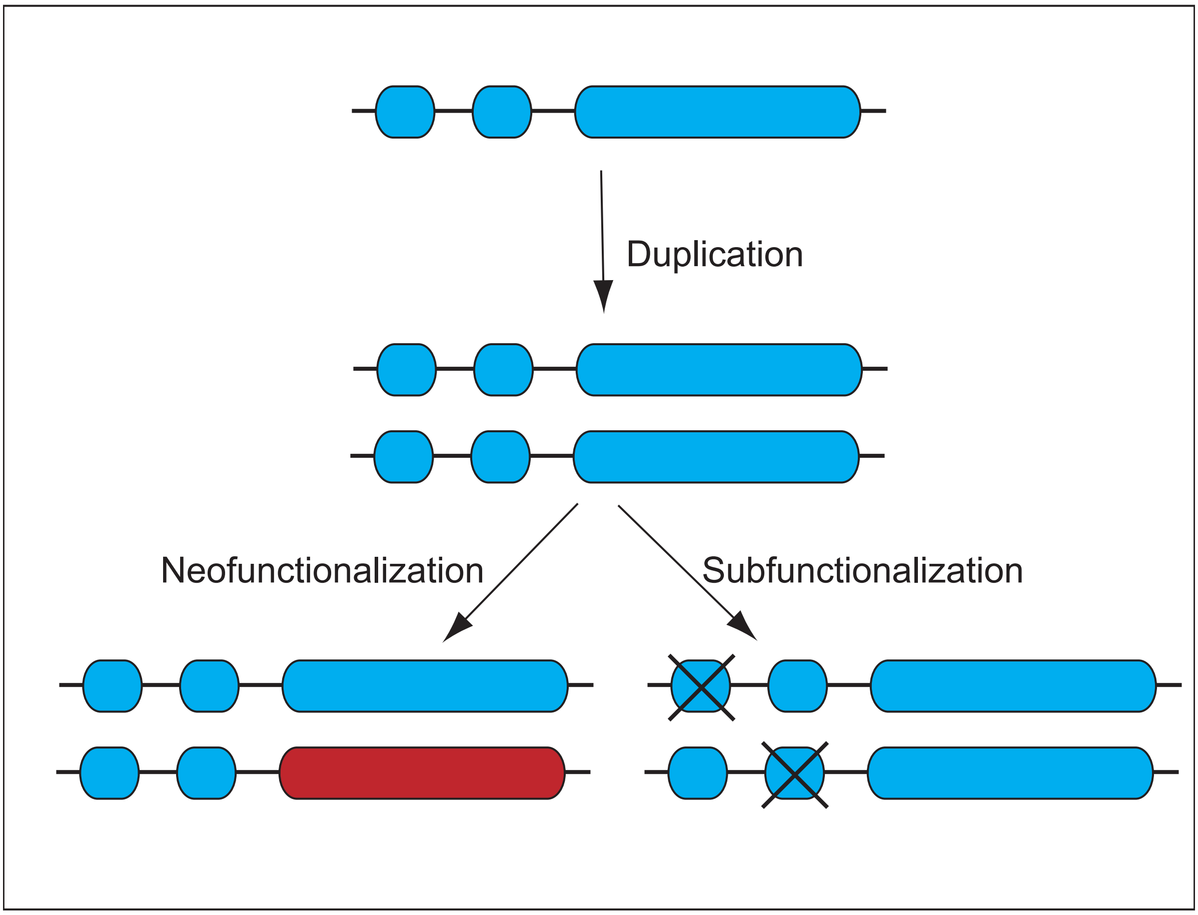


Figure 1.1: Sub- and Neofunctionalization.
A gene with two cis-regulatory modules (short bars), conferring expression in two different tissues, and a single open reading frame (long bar) is duplicated and then either neofunctionalized, where one copy evolves a new (red) function, one retains the original function, or subfunctionalized, where each copy loses a subfunction to degenerative (X) mutations.

1.3 Orthologs and Paralogs

The original definition of orthologs is two genes from two different species that derive from a single gene in the last common ancestor of the species. Paralogs are defined as genes that derive from a single gene that was duplicated within a genome. The latter definition does not specify that paralogs can only be found in a single organism, and hence genes in different organisms that arose from gene duplication in an ancestral genome are also paralogs according to the definition.

Several other aspects of orthologous and paralogous relationships between genes have emerged as important in evolutionary genomics. Figure 1.2 illustrates how multiple genes can simultaneously be orthologs of another gene, in this case HA* can be said to be 'co-orthologs' of WA* (where HA* indicates all genes whose name starts with HA, etc.) Co-orthologs are thus paralogs produced by duplications of orthologs subsequent to a given speciation event (also called lineage-specific expansions of paralogous families), which is commonly observed between distantly related species (Jordan *et al.*, 2001; Remm *et al.*, 2001; Lespinet *et al.*, 2002). This special type of paralog needs a qualifier to distinguish it from paralogs that resulted from an ancestral (relative to the given speciation event) duplication and, consequently, are not (co)orthologous to a given gene in the second species (e.g. HA* and WB in figure 1.2).

Out-paralogs and in-paralogs are derived by analogy to terms used in phylogenetics, 'outgroup' and 'ingroup', which denote anciently and recently branching lineages, respectively. Relative to a given speciation event, paralogs derive either from an ancestral duplication and do not form orthologous relationships, or they derive from a lineage-specific duplication, giving rise to co-orthologous relationships. The logical terms therefore seem to be, respectively, 'out-paralog' and 'in-paralog', explicitly denoting that they are subtypes of paralogs and when they branched relative to the given speciation event.

Therefore, definition of 'in-paralogs' is: paralogs in a given lineage that all evolved by

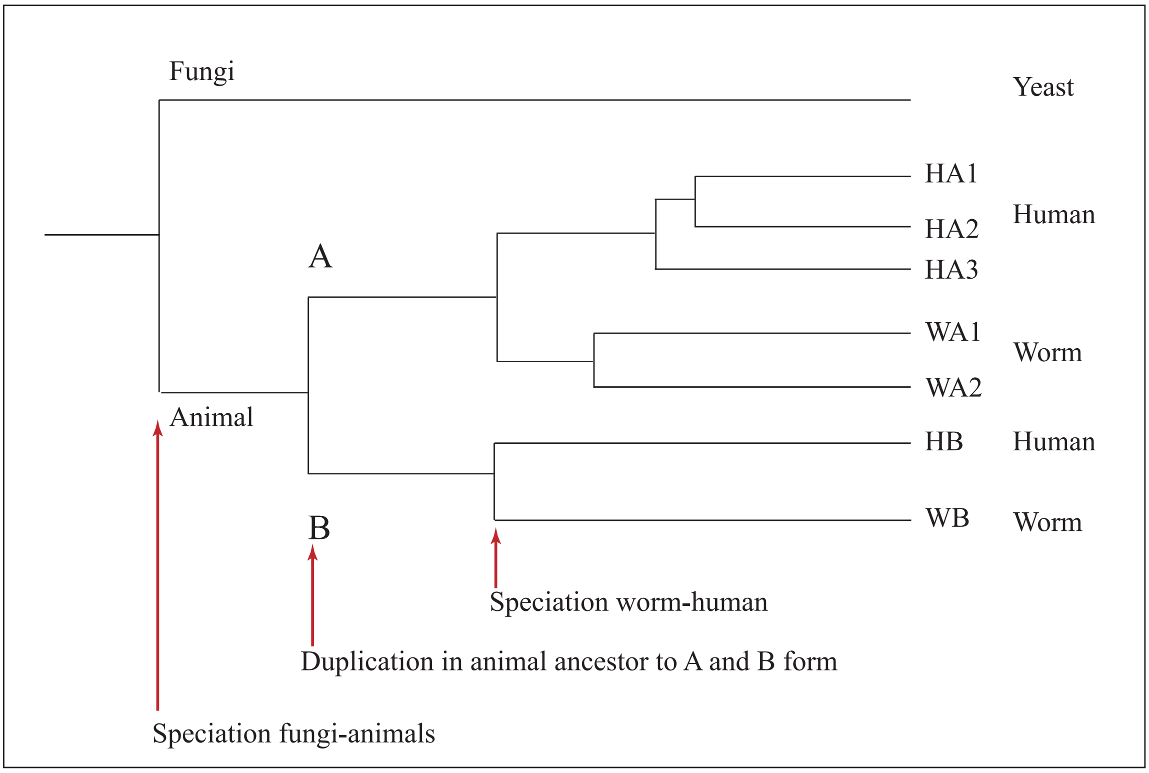


Figure 1.2: In- and out-paralogs.
Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human-worm split, into genes A and B. After the human-worm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'in-paralogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'out-paralogs' when comparing human with worm. However, HB and HA*, and WB and WA* are in-paralogs when comparing with yeast, because the animal-yeast split pre-dates the HA*-HB duplication. (modified from (Sonnhammer and Koonin, 2002))

gene duplications that happened after the radiation (speciation) event that separated the given lineage from the other lineage under consideration. Definition of 'out-paralogs' is: paralogs in the given lineage that evolved by gene duplications that happened before the radiation (speciation) event.

1.4 In-paralogs

In-paralogs by definition are taxon specific. Thus comparing any 2 species and finding their respective in-paralogs I identify all the new genes which has arisen since these species speciation. In-paralogs give us insight in what happened to the both organism since their splitting. Comparing human and chimp would reflect their changes respective to each other in the last 5 million years, comparing human and and mouse - in the last 50, etc. Any two species could be analysed in this way.

The in-paralogs could be analysed individually. So single genes responsible for some specific characteristic could be identified. Grouping in-paralogs functionally would show which functions were under more or less evolutionally pressure. Of course, different functional classifications could be applied (GO classification (Ashburner *et al.*, 2000; Harris *et al.*, 2004), KEGG database (Kanehisa *et al.*, 2004, 2006, 2008)). Also in-paralogs could be simply counted, reflecting the gene duplications rates in the respective time frame.

2 Methods

2.1 Data

For quantitative analysis genomes of following species were used: *Aedes aegypti* (AaegL1) (Nene *et al.*, 2007), *Anopheles gambiae* (AgamP3) (Holt *et al.*, 2002), *Bos taurus* (Btau_2.0) (Snelling *et al.*, 2007), *Caenorhabditis elegans* (CEL160) (Ainscough *et al.*, 1998), *Canis familiaris* (BROAD D1) (Lindblad-Toh *et al.*, 2005), *Ciona intestinalis* (JGI 2) (Dehal *et al.*, 2002), *Ciona savignyi* (CSAV 2.0) (Small *et al.*, 2007), *Danio rerio* (ZFISH6), *Drosophila melanogaster* (BDGP4.3) (Adams *et al.*, 2000), *Gallus gallus* (WASHUC1) (Hillier *et al.*, 2004), *Gasterosteus aculeatus* (BROAD S1) (Gasterosteus aculeatus. Fishbase), *Homo sapiens* (NCBI36) (Lander *et al.*, 2004), *Macaca mulatta* (MMUL_1.0) (Gibbs *et al.*, 2007), *Monodelphis domestica* (BROAD O3) (Mikkelsen *et al.*, 2007), *Mus musculus* (NCBIM36) (Waterston *et al.*, 2002), *Oryzias latipes* (MEDAKA1) (Kasahara *et al.*, 2007), *Pan troglodytes* (CHIMP2.1) (Mikkelsen *et al.*, 2005), *Rattus norvegicus* (RGSC3.4) (Gibbs *et al.*, 2004), *Takifugu rubripes* (FUGU4) (Aparicio *et al.*, 2002), *Tetraodon nigroviridis* (TETRAODON7) (Jaillon *et al.*, 2004), *Xenopus tropicalis* (JGI4.1) (Morin *et al.*, 2006)(Bowes *et al.*, 2008).

For qualitative analysis of following species were used: *Anopheles gambiae* (MOZ2a) (Holt *et al.*, 2002), *Apis mellifera* (Apis 2.0) (Weinstock *et al.*, 2006), *Drosophila melanogaster* (BDGP4) (Adams *et al.*, 2000). All peptides sequences were obtained from www.ensembl.org (Birney *et al.*, 2006)(Flicek *et al.*, 2008).

2.2 Timing the origin of paralogs

Following software packages were used for data analyses:

- Inparanoid (in-paralogs search) (Remm *et al.*, 2001)
- BLAST (Altschul *et al.*, 1997)
- R (statistics) (R Development Core Team, 2008)
- SplitsTree4 and njplot (trees representation) (Perriere and Gouy, 1996)(Huson and Bryant, 2006)

For statistical analysis Fisher's exact test and Pearson's chi-squared test were used. The genomes were analysed pairwise, all vs. all.

2.2.1 Inparanoid

Inparanoid algorithm begins with detection of orthologs, based on calculation of pairwise similarity scores between all sequences. The idea is that if the sequences are orthologs, they should score higher with each other than with any other sequence in the other genome.

As input, program expects two datasets of protein sequences in FASTA format. The datasets should be in two different files and are expected to include the complete set of protein sequences from two species. There could be also a third dataset, so called outgroup. On the evolution tree the outgroup should outside any branch with two analysed species. The potential ortholog pair is deleted if pairwise score is lower than their score against any outgroup sequence.

The detection of orthologs starts with calculation of all pairwise similarity scores between all studied sequences. This is usually done with the BLAST program for speed, but it could be done with any other pairwise alignment program. For datasets A and B, the similarity scores are calculated in four different steps: A vs. A, A vs. B, B vs. A and B vs. B. If there was out-group dataset C additionally similarity scores are calculated for A vs. C and B vs. C.

The clustering algorithm detects non-overlapping groups of orthologous sequences using pairwise similarity scores obtained in the BLAST step. As scoring matrix I used BLOSUM62. However, for distant related and close related species, respectively BLOSUM45 and BLOSUM80 are preferred. Also there is option to use PAM30 and PAM70.

Two adjustable cut-off values are applied to each pairwise match: a score cut-off and an overlap cut-off.

The score cut-off is necessary to separate significant scores from spurious matches. We used score cut-off of 50 bits. The effect of this cut-off is mainly to avoid inclusion of insignificant hits and thereby reduce the volume of data.

Although BLAST is fast and detects biologically relevant homologies reliably, it should be used with caution. The main problem for the presented ortholog detection algorithm is that BLAST reports local similarities. The orthologs are expected to maintain homology over the entire length, or at least over the majority of their length. To avoid domain-level matches, the matched area is forced to be longer than 50% of the longer sequence. This should avoid clustering sequences that share only short domains. For this case the overlap cut-off is applied. As mentioned above overlap cut-off is 50%, i.e. the matching segment of the longer sequence must exceed 50% of its total length.

Thus Inparanoid starts by finding the mutually best hits between species A and B, forming clusters of orthologs. This pair is called ”main ortholog pair of a given ortholog group”. Next, new orthologs are added to clusters if the similarity score between them and the ortholog from the same species from the given cluster is less than similarity score between main ortholog pair from the same cluster (fig. 2.1 and 2.2). As the result, list of orthologs clusters is formed, where clusters represent groups of in-paralogs. Because

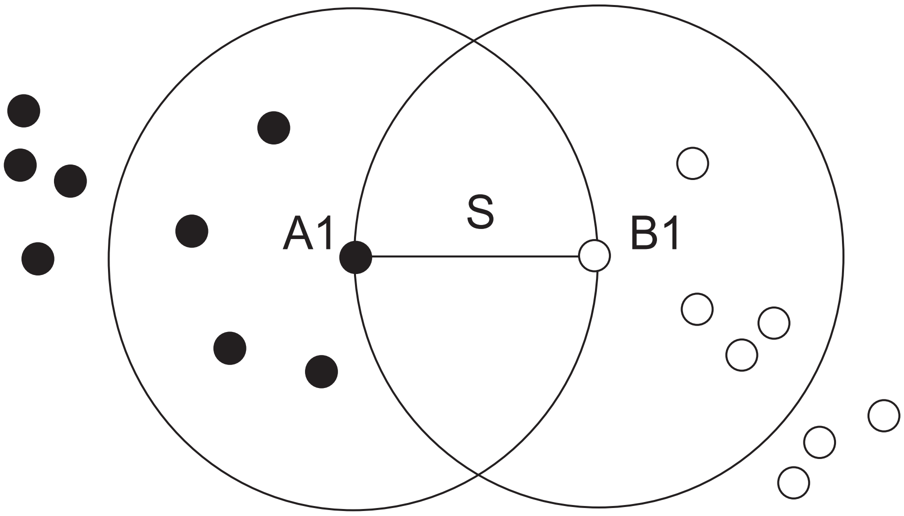


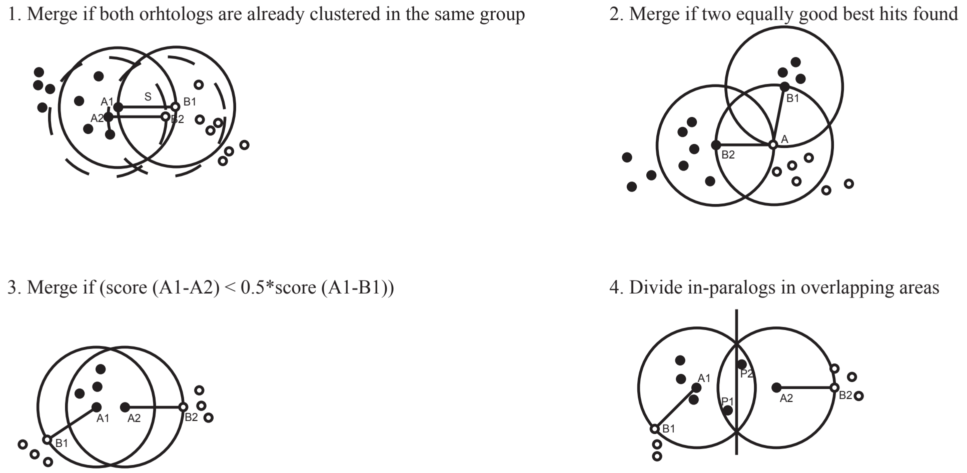
Figure 2.1: Clustering of in-paralogs.
Each circle represents a sequence from species A (black) or species B (grey). Main orthologs (pairs with mutually best hit) are denoted A1 and B1. Their similarity score is shown as S. The score should be thought of as reverse distance between A1 and B1, higher score corresponding to shorter distance. The main assumption for clustering of in-paralogs is that the main ortholog is more similar to in-paralogs from the same species than to any sequence from other species. On this graph it means that all in-paralogs with score S or better to the main ortholog are inside the circle with diameter S that is drawn around the main ortholog. Sequences outside the circle are classified as out-paralogs. In-paralogs from both species A and B are clustered independently. Modified from Remm *et al.* (2001).

in-paralogs were represented by proteins, I mapped them all back to the corresponding genes and removed redundant ones from the clusters (fig. 2.3 and 2.4).

2.2.2 Gene duplications matrix

Then, gene duplication events were counted. For example, if the cluster contained 2 genes from species A - it was counted as 1 duplication event in species A, 3 genes - 2 duplication events, 4 genes - 3 duplication events etc.

A matrix, representing all 21 species, showing the number of genes duplications in every species compared to every species was build. Following, the duplications in human



1. Merge if both orhtologs are already clustered in the same group

2. Merge if two equally good best hits found

3. Merge if (score (A1-A2) < 0.5*score (A1-B1))

4. Divide in-paralogs in overlapping areas

Figure 2.2: The rules for resolving overlapping groups of in-paralogs. In-paralogs are clustered in order of their similarity scores, starting with the more similar groups. The rules are applied in the following order: (1) merge groups if main orthologs A2 and B2 are already clustered in the same group with a stronger group A1-B1; (2) merge groups if main ortholog A has equally best hit to two orthologs from B, B1 and B2; (3) merge groups if one of the new ortholog candidates already has a high confidence value in another group; (4) all other overlapping groups of in-paralogs are separated based on their distance to the main ortholog. In the given example, the in-paralog P1 will remain in group with A1, but the in-paralog P2 will be moved into the second group with A2. Modified from (Remm *et al.*, 2001).

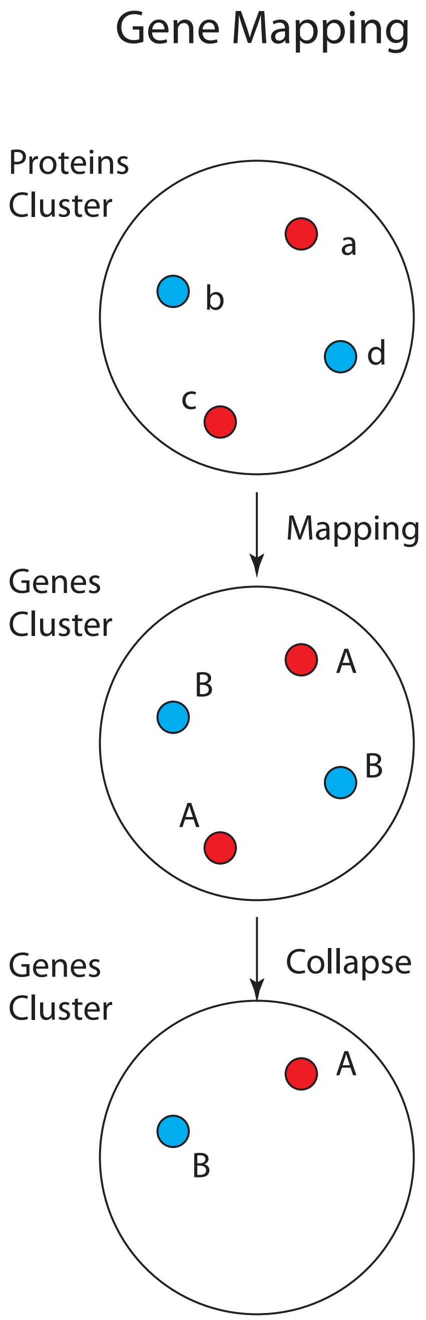


Figure 2.3: Gene mapping.
After Inparanoid step as result I get the list of protein clusters containing in-paralogs. Then every protein is mapped to its gene. And as the last step all the redundant genes are removed.

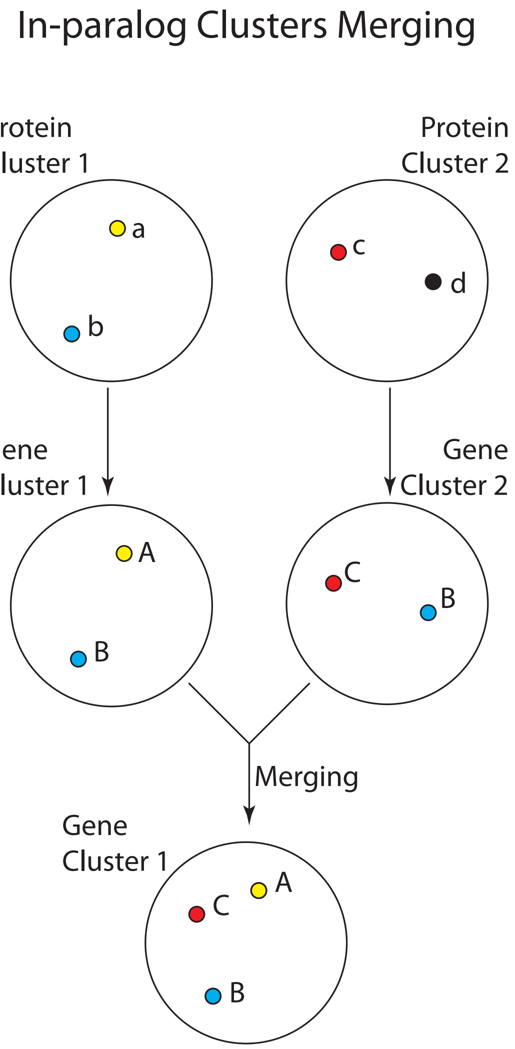


Figure 2.4: In-paralog clusters merging.
After gene mapping step I have to merge all the gene clusters containing the same genes.

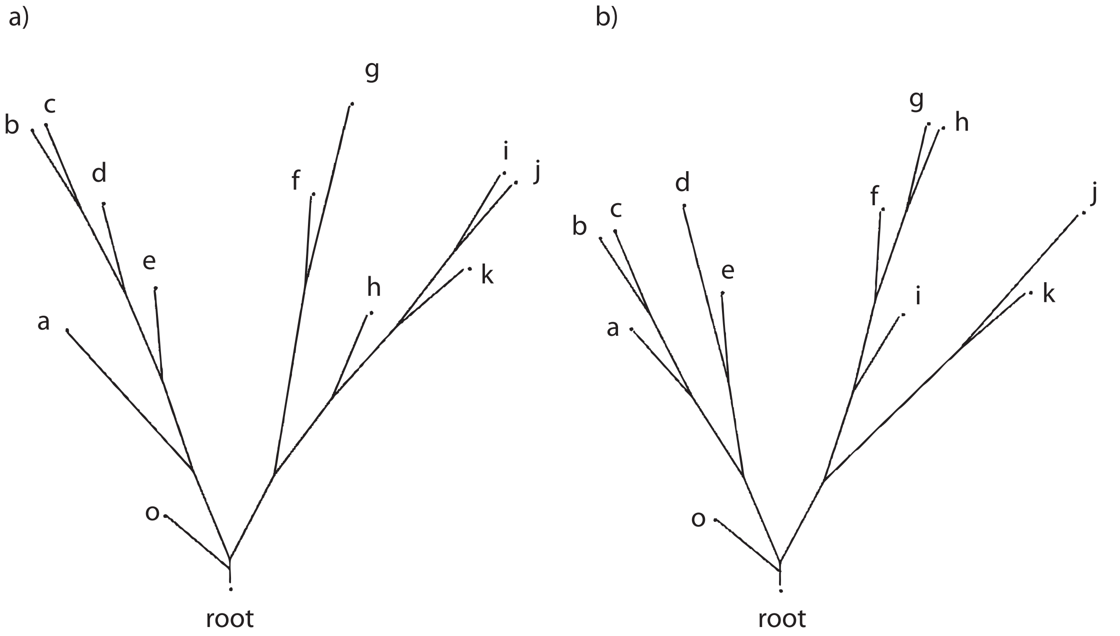


Figure 2.5: Two different rooted phylogenetic tree.

compared to chimp could be timed as the ones which happened in the last 5 MYR, between human and mouse - 90 MYR etc.

2.2.3 SplitsTree4

A "phylogenetic tree" is commonly defined as a leaf-labeled tree that represents the evolutionary history of a set of taxa, possibly with branch lengths, either unrooted or rooted.

However phylogenetic network is more complicated term is defined as "any" network in which taxa are represented by nodes and their evolutionary relationships are represented by edges. (For phylogenetic trees, edges are referred to as branches.) Under this very general heading, one can distinguish between a number of different types of networks. Phylogenetic trees constitute one type (fig. 2.5).

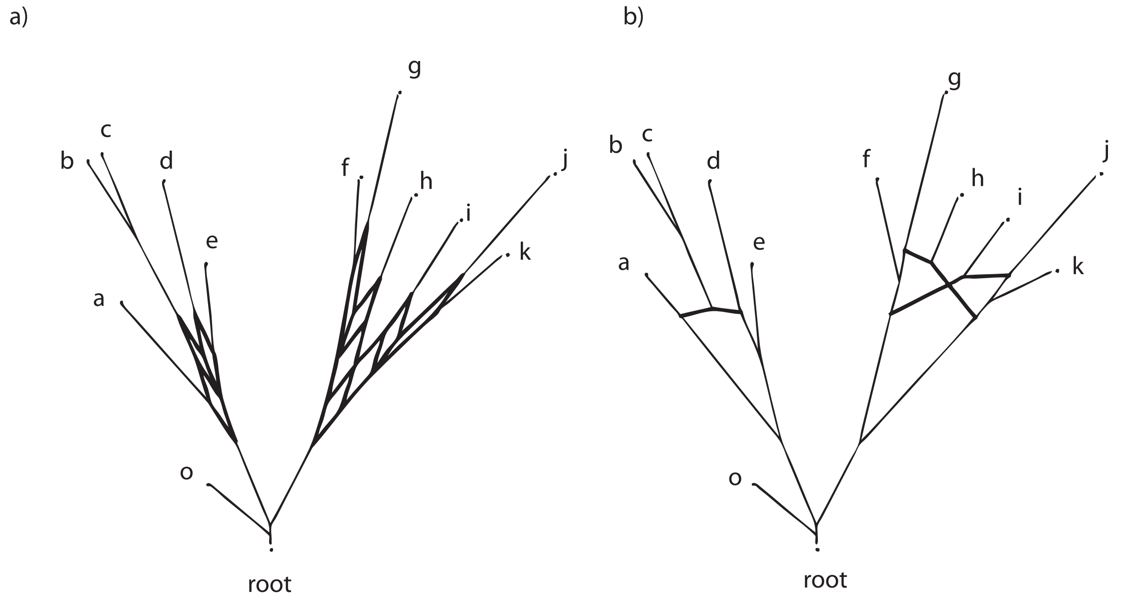


Figure 2.6: Phylogenetic Networks.
(a) A split network representing all splits present in the two trees depicted in the previous figure. Here, each band of parallel edges corresponds to a branch contained in one of the input trees. The nodes do not necessarily correspond to hypothetical ancestors. (b) A reticulate network that explains the two trees by postulating three reticulations that give rise to the clades (b, c), (h), and (i). This network explicitly describes a putative evolutionary history: the internal nodes correspond to ancestral taxa, and the edges represent patterns of descent.

A second type is the "split network," which is obtained as a combinatorial generalization of phylogenetic trees and is designed to represent incompatibilities within and between data sets (fig. 2.6a). A third type, "reticulate network," represents evolutionary histories in the presence of reticulate events such as hybridization, horizontal gene transfer, or recombination (fig. 2.6b).

Reticulate networks provide an "explicit" representation of evolutionary history, generally depicted as a phylogenetic tree with additional edges. The internal nodes in such a network represent ancestral species, and nodes with more than two parents correspond to reticulate events such as hybridization or recombination.

Split networks are used to represent incompatible and ambiguous signals in a data set. In such a network, parallel edges, rather than single branches, are used to represent the splits computed from the data. To be able to accommodate incompatible splits, it is often necessary that a split network contains nodes that do not represent ancestral species. Thus, split networks provide only an ”implicit” representation of evolutionary history.

There are many algorithms for inferring the split networks. Split decomposition (Bandelt and Dress, 1992) and neighbor-net (Bryant and Moulton, 2004) construct split networks based on given distance matrices. In SplitsTree4 software both of these methods are implemented. For our dataset I used neighbour-net algorithm.

2.2.4 Simulation of gene duplications

Gene duplications were simulated for all 21 species mentioned above. For the simulation the evolutionary distances were assigned as in figure 2.7. As a common ancestor for all species I used a genome containing 1,000 genes. One step of the simulation process was equal 1 MYR. Thus, from the common ancestor till present time 1,177 simulation steps were performed for every species. The duplication rate was calculated as a function of time and was calculated anew for every step of simulation and assigned for a whole genome and not for a single gene. If the duplication rate was 0.01 it means that in whole genome consisting of 1,000 genes approximately 10 would duplicate. At every speciation point the number of new duplicates was counted. Thereby I knew the new genes number for every single speciation time point in respect to previous one.

2.2.5 Programming

All the programming steps (building the gene duplication matrix, simulation of gene duplication) were done using Ruby (Thomas *et al.*, 2005).

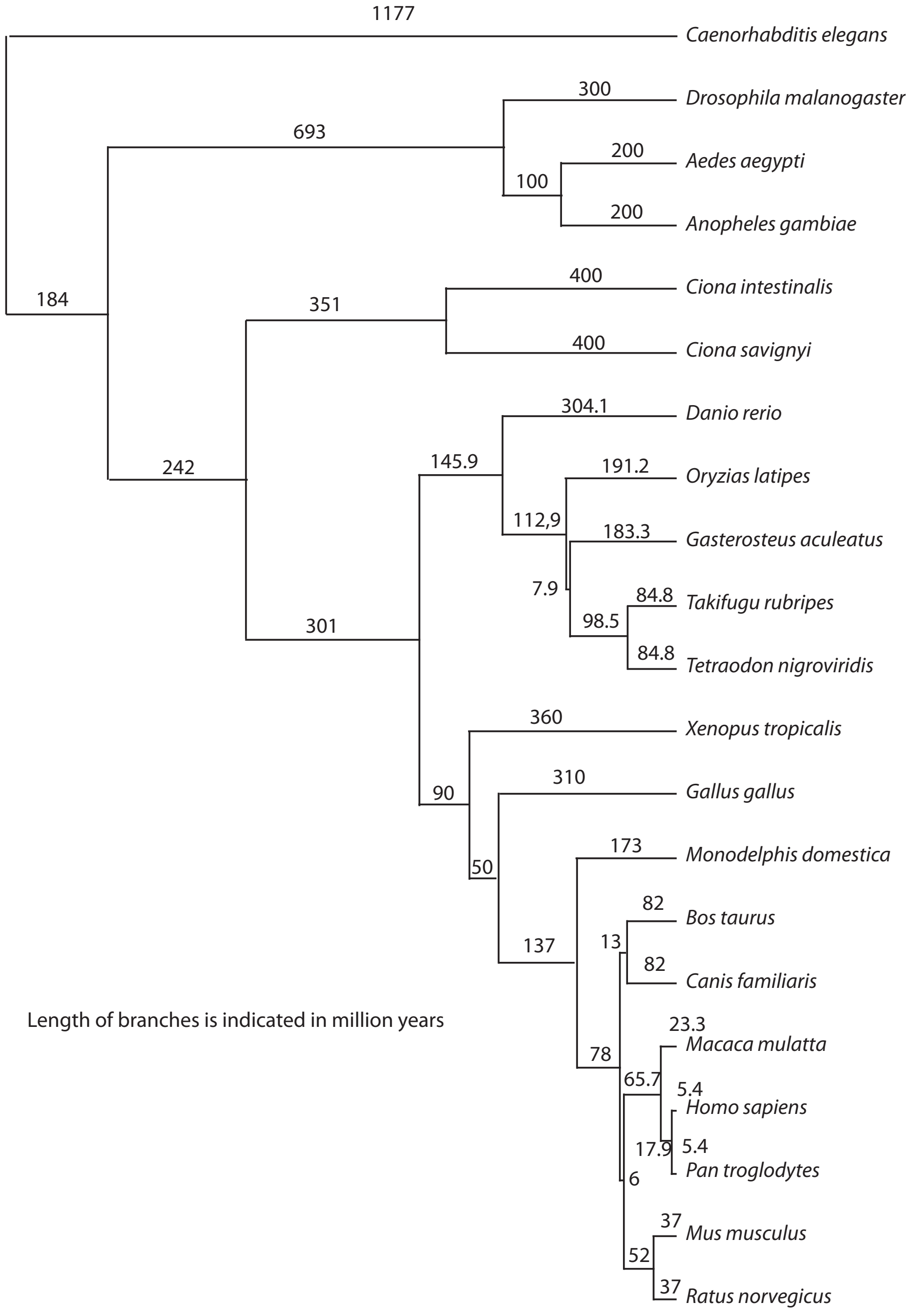


Figure 2.7: Time Tree.
Phylogeny tree of 21 species based on molecular clock and fossil data (Blair Hedges and Kumar, 2003). For 2 speciation events (between Ciona intestinalis and Ciona savignyi, and between Aedes aegypti and the rest of insects) there was no fossil records and molecular analyses available, therefore the speciation events was estimated using the actin protein for molecular clock analysis.

3 Quantitative analysis

3.1 Introduction

An analysis of gene duplications and losses in humans rejected a constant-rate birth death process when looking at larger time scales (Cotton and Page, 2005). This analysis was based on the topology of selected gene tree families. Contrasting, an analysis of 12 *Drosophila* genomes focussed on all orthologs and paralogs within these genomes. Again, a skew in the rate of duplication was found, as most arose in recent events (Heger and Ponting, 2007).

Here I studied duplication patterns but in broader range of species and bigger time scale. We have chosen 21 species, with available fully sequenced genomes. The analysed species cover insects, fish, and various mammals, with evolutionary distances between them from 5 up to 1,000 million years (MYR) and thus in-paralogs ranged from 5 to 1,000 MYR old. For each pair of species I identified species specific duplications. By integrating these data with a literature based phylogenetic tree I could estimate the age of duplications.

	Homo Sapiens	*Mus musculus*
Cluster 1	ENSG00000204435 ENSG00000206300 ENSG00000206406	ENSMUSG00000024387
Cluster 2	ENSG00000066136	ENSMUSG00000032897 ENSMUSG00000073233
Cluster 3	ENSG00000107643 ENSG00000109339	ENSMUSG00000021936 ENSMUSG00000046709
Cluster 4	ENSG00000099917	ENSMUSG00000012114
Cluster 5	ENSG00000135486 ENSG00000139675 ENSG00000176757 ENSG00000187999	ENSMUSG00000046434 ENSMUSG00000058922
Etc...	...	...

Table 3.1: Clusters of in-paralogs between Homo sapiens and Mus musculus. The table represents actual gene clusters form analysis human and mouse genomes. Clusters one contains 2 duplications in human, cluster two - one duplication in mouse, cluster three - one duplication in each of the species, cluster four - no duplications, and cluster five - three duplications in human and one in mouse. Complete list has 15135 clusters.

3.2 Results and Discussion

3.2.1 A species tree based on the number of accepted in-paralogs

To identify gene duplications which arose after species split (in-paralogs), proteomes of 21 different species were pairwise analysed by the Inparanoid algorithm (Remm *et al.*, 2001). Following, each protein was mapped onto its gene. Clusters, containing only single genes as result of mapping, were omitted. The remaining contained multiple genes for one or both species, indicating a species specific expansion (tab. 3.1).

Based on the clusters with multiple genes the number of lineage specific gene duplications was counted for each pairwise comparison (see table 3.1 and 3.2).

For example in human the duplications number increases with increasing of evolutionary distance (the lowest - 1,251 compared to chimp and the highest - 2,814 compared to *Drosophila*). However the number of duplication which happened in chimp compared

	Danio rerio	*Drosophila melanogaster*	*Homo sapiens*	*Mus musculus*	*Pan troglodytes*	*Xenopus tropicalis*
Danio rerio	0	3559	4489	4187	4301	4345
Drosophila melanogaster	527	0	584	552	557	577
Homo Sapiens	2529	2814	0	1992	1251	2469
Mus musculus	2259	2883	1963	0	1783	2404
Pan troglodytes	2096	2485	504	1341	0	1989
Xenopus tropicalis	2017	2166	1846	1929	1879	0

Table 3.2: Lineage specific duplications of selected species (in-paralogs matrix). The table shows the number lineage specific duplications after the speciation event took place. The name of the row indicates duplicated genes' species and name of the column indicates species to which comparison was made. Because of the size constraints I show here only 6 species. The complete version of the table with all 21 species could be found at http://domains.bioapps.biozentrum.uni-wuerzburg.de.

to human (504) is more than 2 times lower than the corresponding number for human (1,251). It indicates that the speed of duplication (duplication per MYR) could be very different. To analyse whether the number of in-paralogs and thereby of accepted duplications was correlated with the species phylogeny, I calculated a phylogenetic tree based on the numbers of in-paralogs. Therefore I symmetrised the in-paralogs matrix by adding the corresponding values for both directions of the pairwise analysis (tab. 3.3).

In-paralogs could be seen as evolutionary distances between species. Thus table 3.3 is a distance matrix. To symmetrise matrix including 21 species "Ward" clustering algorithm was applied (fig. 3.1).

Based on the duplications phylogeny tree could not be properly reconstructed. At least for a large number of species. Only when the species with relatively equally evolutionary rate are compared it could result in tree similar to the real one (fig. 3.2).

Another way of graphically representing the distance matrix is neighbour-net method

	Danio rerio	*Drosophila melanogaster*	*Homo sapiens*	*Mus musculus*	*Pan troglodytes*	*Xenopus tropicalis*
Danio rerio	0	4086	7018	6446	6397	6362
Drosophila melanogaster	4086	0	3398	3435	3042	2743
Homo Sapiens	7018	3398	0	3955	1755	4315
Mus musculus	6446	3435	3955	0	3124	3918
Pan troglodytes	6397	3042	1755	3124	0	3868
Xenopus tropicalis	6362	2743	4315	3918	3868	0

Table 3.3: Symmetrised matrix of lineage specific duplications of selected species (inparalogs).
The table 3 represents symmetrised table 2. Each value was calculated by adding the corresponding values from each pair of species. For example, 7018 in *Homo sapiens* and *Danio rerio* cell were calculated by adding the duplications in *Homo sapiens* to *Danio rerio* (2529) and duplications in *Danio rerio* to *Homo sapiens* (4489). In this way the matrix became symmetrised and clustering for building phylogeny tree could be applied.

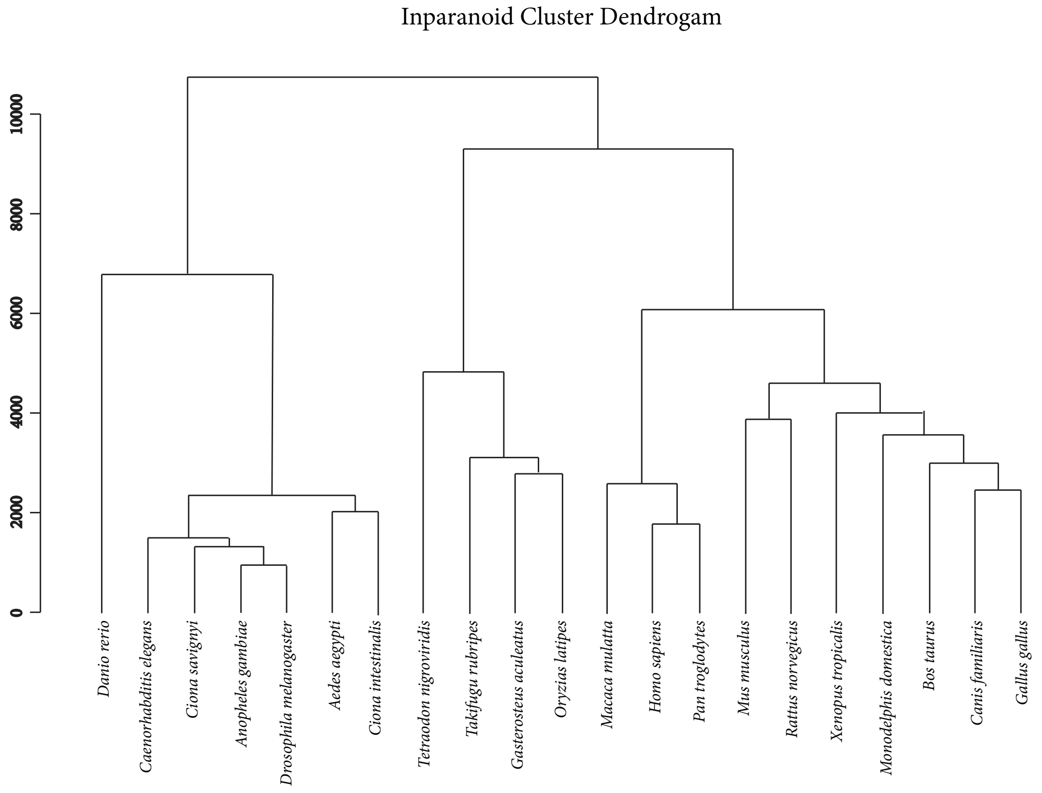


Figure 3.1: Phylogeny tree based on in-paralogs in 21 species.

In figure 10 phylogeny tree for all 21 species from our analysis is shown. As seen a lot of species are placed in wrong cluster (*Danio rerio*). *Caenorrhabditis elegans* is not represented as separate leaf and grouped together with insects and cionas. Also inside mammals group positions of a lot of species is wrong.

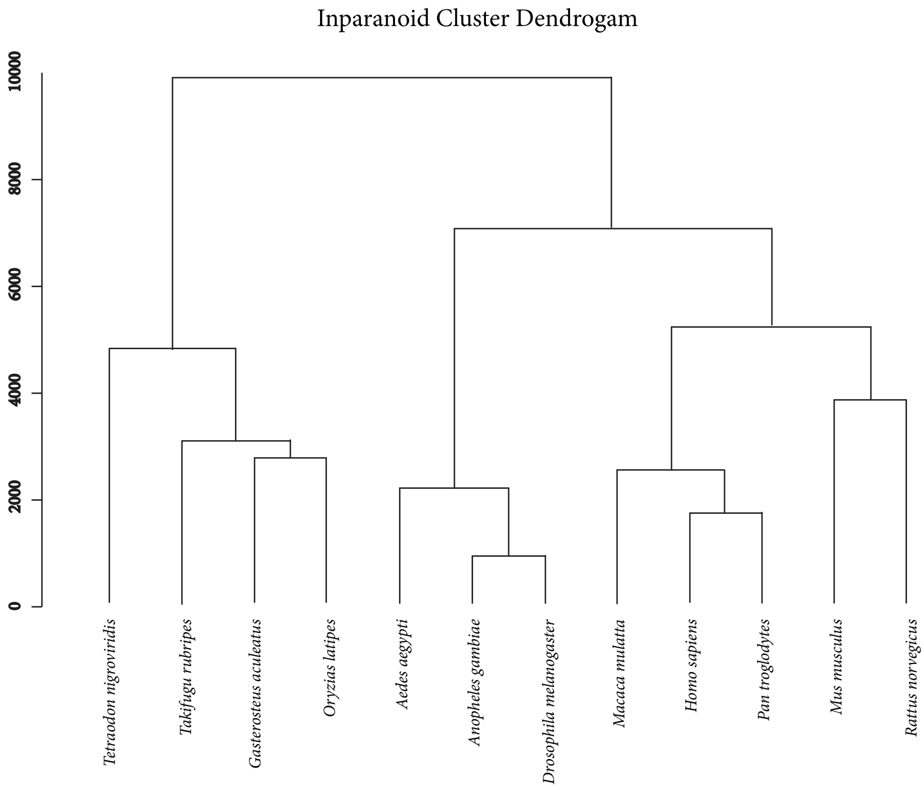


Figure 3.2: Phylogeny tree based on in-paralogs in 12 species.
We reduced the number of species to12, removing all the species with too fast or too slow duplications rate. After that all the species are placed correctly, except that insect group is more closely related to mammals than fish.

(Huson and Bryant, 2006) (3.3). This method allows more than one possible branching. And all the possibilities are represented. Thus I can see all the alternative species splittings.

On many branches the tree followed the standard phylogeny. For example, all mammals are grouped together. A case of unclear placement is the chosen insects, namely *Drosophila melanogaster*, *Aedes aegypti* and *Anopheles gambiae*. Thus, the rate of in-paralogs differs in this taxon from the general rule. This could be caused either by a decreased invention or an increased loss. As an increased loss of orthologs has been shown for *Hymenoptera* the second is likely (Wyder *et al.*, 2007). Similarly, the short branch of the chicken could be caused by the poor quality of the genome assembly (Hillier *et al.*, 2004). Thus, with reasonable exceptions, the rate of accepted in-paralogs reflects phylogeny.

3.2.1.1 Rate of accepted in-paralogs

As the next step, I aimed to correlate the results with the assumed time of divergence between the analysed species. We therefore manually generated a species tree with the length of the branches according to published times of divergence, based on fossils and molecular clock data (fig. 2.7). By dividing the number of in-paralogs by the length of the branch I got the rate of duplications for each branch of the tree (fig. 3.4). For internal branches, the duplication rate was calculated by dividing the duplication rate for all the species in this branch by the distance from the first speciation event till the present.

This tree indicated an increase of the accepted duplications rate in recent times compared to the past. For example, in humans in the last 5 million years the rate reaches 230 duplications per MYR (0.01 duplications per gene per MYR). To analyse whether this trend is real and to overcome possible inconsistencies within the tree, I compared each single species with all others. Therefore, I plotted the number of in-paralogs against

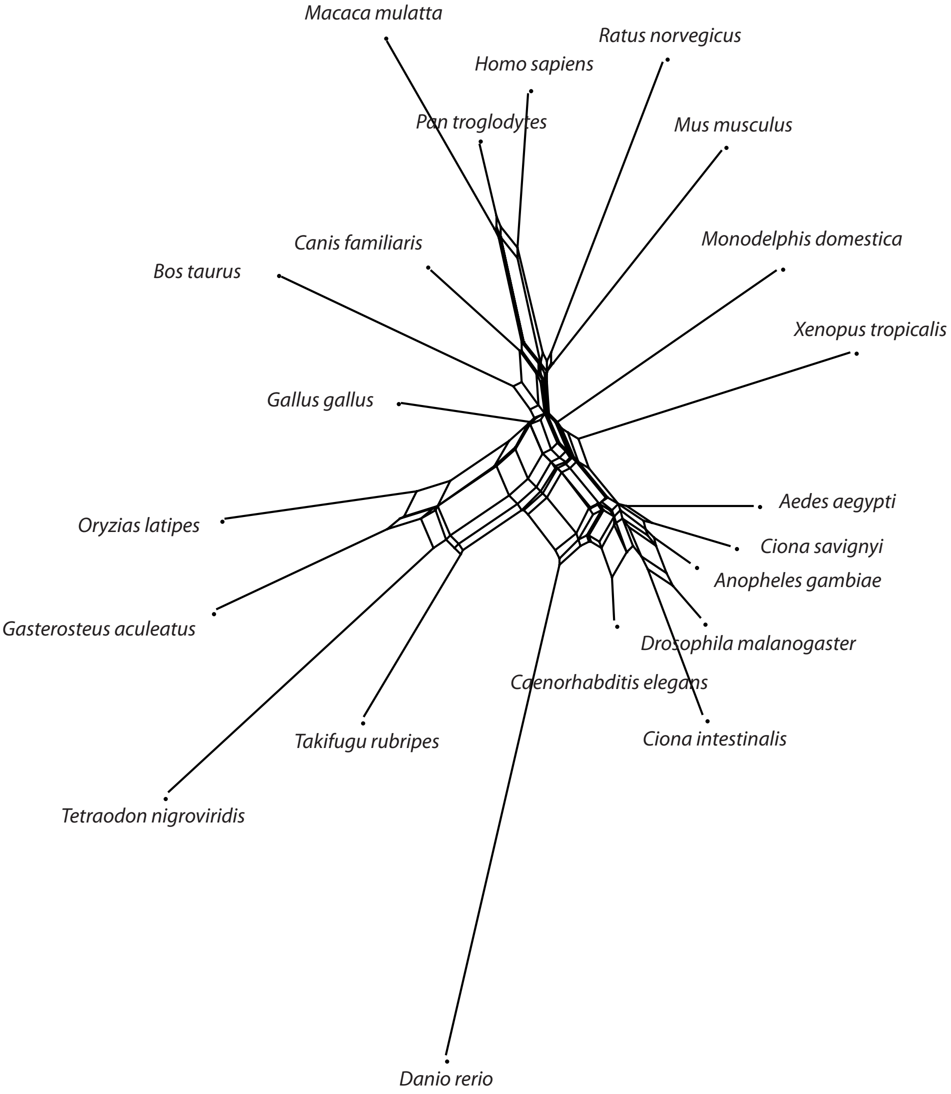


Figure 3.3: Neighbor-Net.
The net includes all 21 species used in the Inparanoid analyses. Using a neighbour-net method the inconsistency in distance matrix data could be visualised. E.g. based on gene duplication data it is impossible to define if chicken is closer related to fish or cow. The length of branches is proportional to number of duplications. Fish, especially *Danio rerio*, *Monodelphis domestica*, *Xenopus tropicalis* and primates have a relatively high rate of gene duplications compared to, for examples, insects.

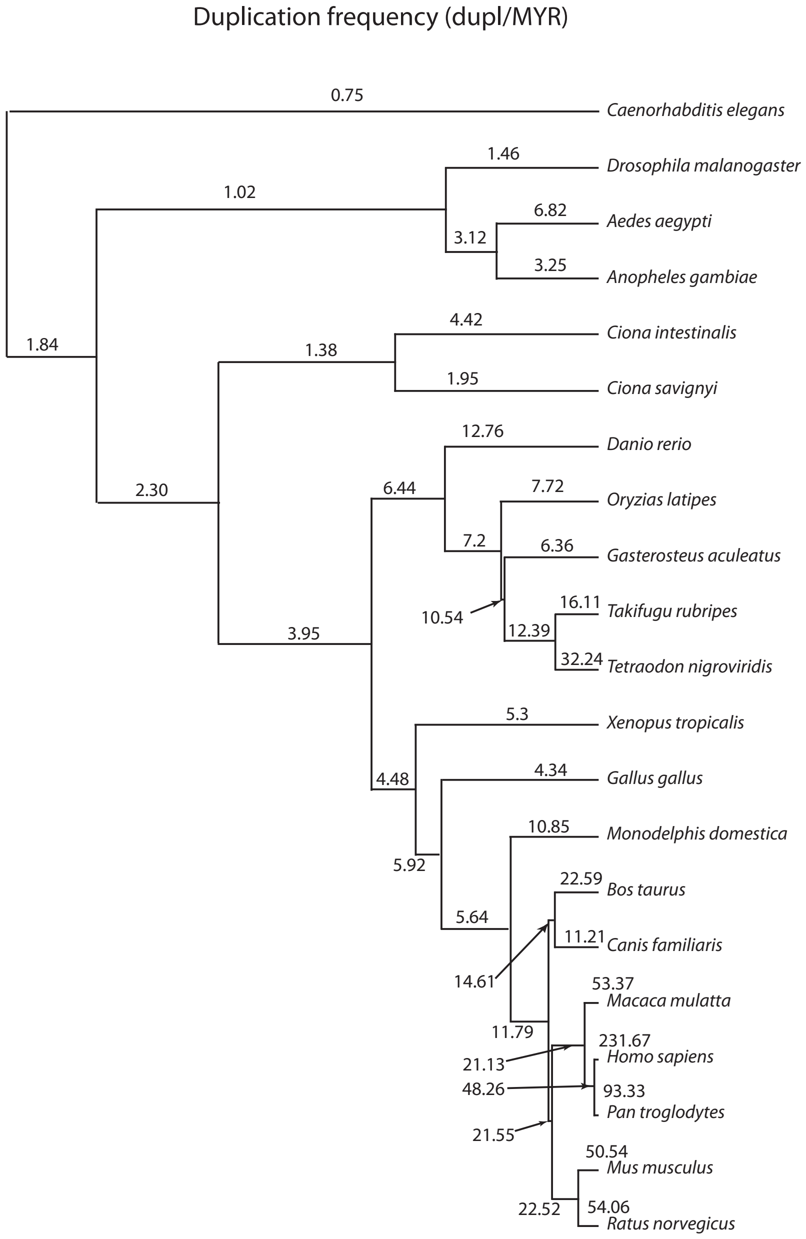


Figure 3.4: Accepted duplications frequency.
The duplication frequency for each branch was calculated by dividing the number of in-paralogs by the length of the branches in million years (MYR). If more than one species were on the branch, the average number of duplications number was taken. When getting closer to the base of the tree, the results are averaged over more species and the duplication frequency characterises a whole taxonomic group. However it is obvious that duplication rates drop significantly closer to the base of the tree, or, in other words, less amount of duplication has survived until present among those who appeared closer to the root of the tree.

the evolutionary distances for each single species. Exemplary, figure 3.5 (experimental data) shows the comparison of *Homo sapiens* to all other species. Indeed, I found that the number of accepted duplications is not increasing linearly over time but saturates and most of the duplicated genes are coming from the near-present time. In addition I simulated the process of gene duplication (fig. 3.5 simulated data). We set the ancestor's genome to 1000 genes. In the simulation process each step was 1 MYR and the duplication rate represented as a function of time. The duplication rate was set close to zero in the beginning of evolution and then it was steadily increasing, reaching its maximum at the present. The evolutionary process was simulated for all 21 species in our study with speciation events according to figure 2.7.

Whereas all mammalian experimental data graphs were qualitatively the same as the human one, fish and insect showed a decrease of the number of duplications when compared to the most divergent species (Fig. 15 experimental data). Thus, I detect less duplications when compared to species distanced 1,000 MYR than compared to the species distanced 750 MYR. For example, in humans the number of duplications was steadily increasing with larger evolutionary distances, although the rate of accepted duplications 1,000 MYR ago was almost zero. Contrasting, in fish approximately 800 MYR ago the rate of accepted duplications became negative. To reconstruct the shape of the experimental data dots distribution in the simulation for fish and insects at a certain point of evolution the rate of accepted duplications had to become negative and had to steadily decrease closer to the base of the evolutionary tree (fig. 3.6 simulated data).

From these data I can see that most of the accepted duplications I observe today arose recently. Thus, a gene which appeared by duplication 500 MYR years ago has a much lower chance to be observed today than genes which appeared 5 MYR years ago. Here, it has to be considered that if the evolutionary distance between two species is

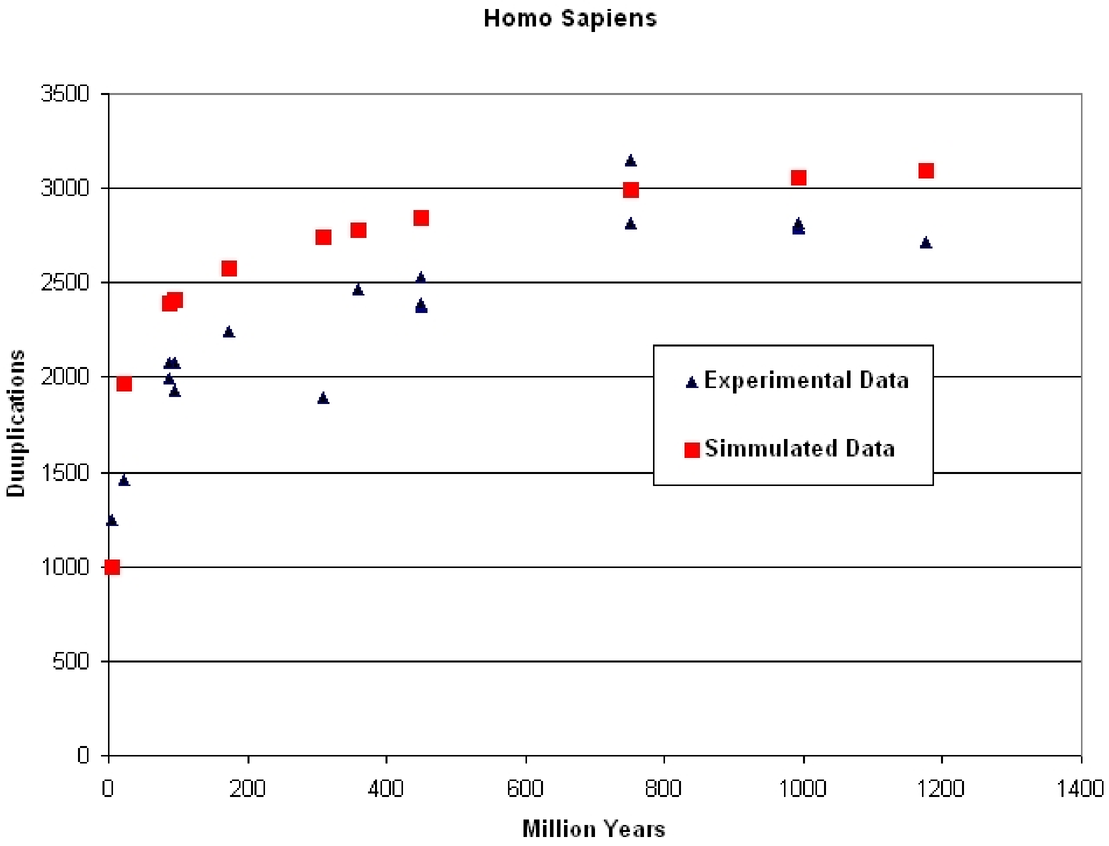


Figure 3.5: Comparison of duplications from *Homo sapiens* to other species.
This figure shows the number of in-paralogs plotted against the time - 0 MYR denotes present and 1,400 MYR is past. Triangles represent experimental, squares - simulated data. In the simulation the duplication rate was set to be high in near-present time (left part of the graph) and decreased gradually in the near-past time (right part of the graph).

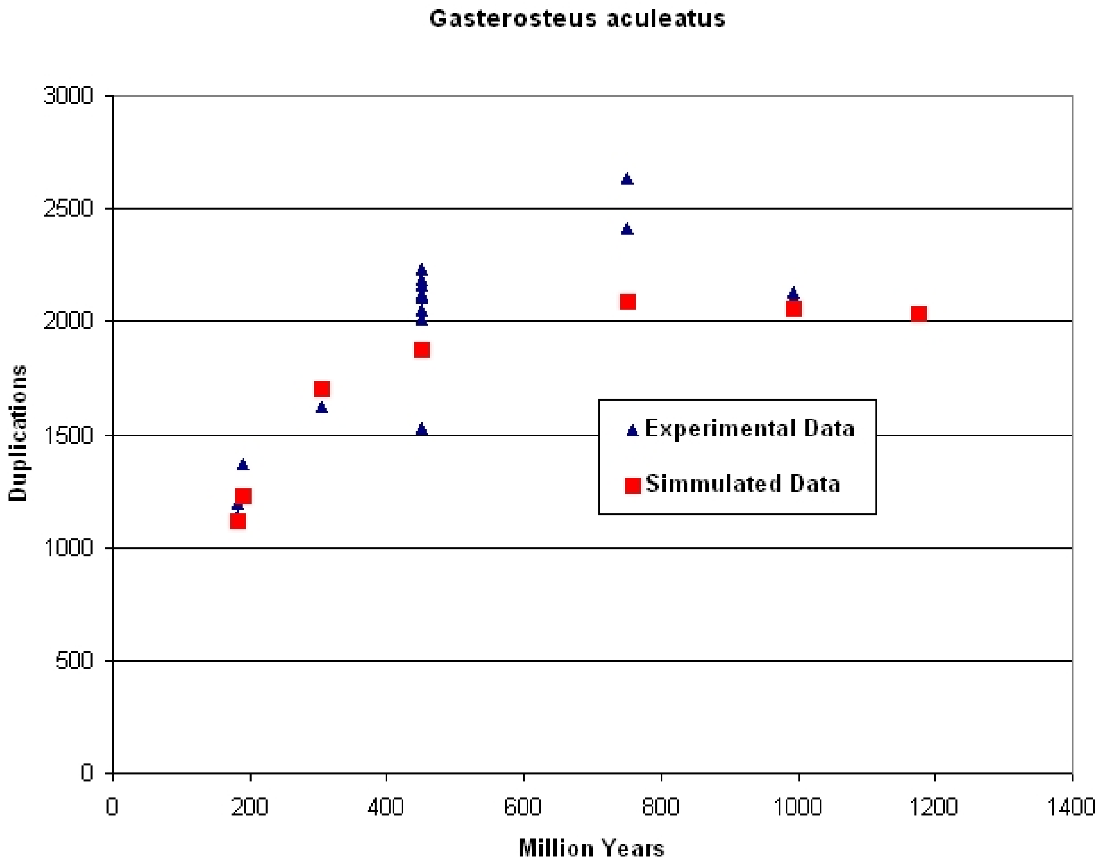


Figure 3.6: Comparison of duplications from *Gasterosteus aculeatus* to other species. On the plot there are represented amount of in-paralogs plotted versus the time. 0 MYR is present and 1400 MYR is past. Triangles represent experimental data, squares represent simulated data.

large enough (more than 800 MYR) the detection of accepted duplications is becoming very difficult. This leads even to a decrease in duplications with increasing evolutionary distance.

3.2.2 Model for duplication rate

The aim of this study was to unravel the history of duplicated genes observed in today's species. In a first step I have built a phylogenetic tree based on the number of pairwise in-paralogs between selected species. Its neighbor-net representation revealed major consistencies, but also incongruence with the tree of life. The latter could be caused by a biological signal like in the case of *Hymenoptera*, but also on low quality of genomic sequences. Additionally, I had to symmetrise the in-paralogs matrix for the neighbor-net analysis. This might have levelled out the biological signal because the in-paralogs distance between any two species was represented as a sum of in-paralogs from both species. In a second step, the point of emergence of an in-paralog observed today was delineated. By correlating this to a timed version of the tree of life, I obtained the number of in-paralogs per million years for each branch (fig. 3.4). This revealed an increase in the rate following more recent speciation events. To analyse this observation in more detail and to reduce the influence of incongruence in the in-paralogs matrix, I looked at the emergence of in-paralogs for different species separately. Here, one has to consider that the species sampling in the analysed tree lead to differences in coverage for different sub-trees. For example in the case of humans, 11 speciation events could be considered, whereas for insects there were only up to 4. These species based analyses further corroborated a difference in the rate of accepted duplications: the further back into the past - the less accepted duplications we observe today.

To describe this effect in more detail, a qualitative simulation was performed. In figure 3.7 I show the accepted duplications rate I used in our simulations. Using this rate I was able to obtain very similar numbers of in-paralogs compared to experimental data. The duplication curves showed a "hollowed-out exponential" shape (Harvey *et al.*, 1994). Thus, genes arisen in 'old' duplications that still exist today can be seen as "survivors". This result is in concordance with recent data on gene in-paralogs over a smaller time scale (between 12 *Drosophila* species) (Heger and Ponting, 2007).

When comparing the species with a high divergence time I observed a decrease in duplications number compared to more closely related ones (fig. 3.6). The reasons for this are possibly limitations of the in-paralog detection combined with species specific features. As a first step of the analyses a Blast search was used to detect the correct ortholog seed. This search has a certain cut-off value. With increasing evolutionary distance sequences of true orthologs diverge further and therefore the alignment score becomes smaller until it is considered to be not significant. Following, it is not treated as ortholog seed by the Inparanoid algorithm and a whole cluster of orthologs would be lost. Accordingly, after reaching certain time of divergence (800 MYR) the number of newly detected in-paralogs is zero or decreasing. This process is very well seen in fish and insects but not that obvious in humans. Probably the rate of proteins evolving in fish and insects was higher than that in mammals. In order to simulate these data the duplication rate had to be set negative.

The number of in-paralogs we observe today is not only shaped by duplications but also by gene loss. Thus duplications observed today are "Duplicated genes" minus "Lost duplicated genes". If one assumes that the gene duplication rate is a stochastic process and the amount of genes in the genome is not changing over time then the duplication rate is constant. Then gene losses are responsible for unequal distribution of duplicated genes in the tree of life.

We can only speculate why in-paralogs observed today tend to have arisen in more recent time. If a species is adapting to a novel ecological niche it will acquire new

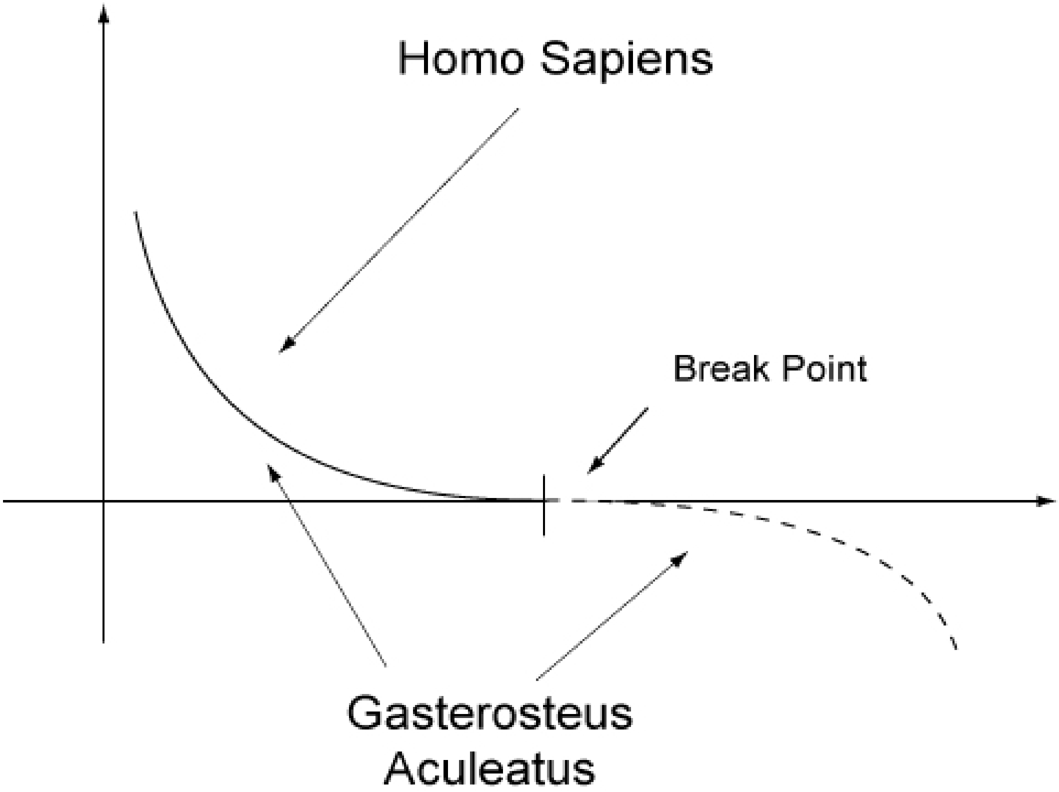


Figure 3.7: X-axis corresponds to time going from present into the past.
For human and other mammal the accepted duplication rates were always positive. It was maximal at the end of evolution and almost zero at the beginning. The accepted duplications rate never reached the "breakpoint" where loss of duplicated genes prevailed over the new duplications. However in fish and insects with the distance between speciation events more than 800 MYR the difference between duplicated genes and loss of duplicated genes begin to decrease more and more. And at certain moment "breakpoint" in accepted duplications rate would be observed.

characteristics, and some of them will be based on duplication of genes. In the same process the evolutionary pressure on genes needed for the adaptation to the previous surrounding will decrease. Thus, the fitness of an organisms having lost one of those genes will not be that strongly affected. Changing ecological niche happened numerous times on the course of last 1,000 MYR of evolution for every species. And every next niche was more and more different in its requirements for species as the original one. Thus the genes needed for survival long time ago have a very slim chance to be present in genomes now.

Such a distribution of accepted gene duplications could be described by a model in which birth and lost rates are always high. Most of the new genes exist only for relatively short period of time with just a few being functional over long periods of time. Such a distribution of accepted duplications is universal and could be seen throughout all the lineages. Taking all together, old genes (which appeared in genome long time ago) are having much higher risk of being lost along the evolution road then the younger ones.

4 Qualitative analysis

4.1 Introduction

In the previous chapter in-parallogs from pairwise comparisons of 21 species were calcualted. In-paralogs were mapped on to the species speciation timescale and duplications rates were derived (fig. 2.7 and 3.4). The next question is the functional classification of in-paralogs. Can we link the functional classes of duplicated genes to morphological features of the species? Because this kind of analyses is much more time consuming I used only three species, namely fruitfly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*) and honeybee (*Apis mellifera*). These are insects with vastly different lifestyles and only insects with fully sequenced genomes at the time of study.

The first speciation event between *Hymenoptera* (*Apis mellifera*) and *Diptera* (*Anopheles gambiae* and *Drosophila melanogaster*) took place in the Triassic or the Upper Permian period 250 million years (Myr) ago, or even 300 Myr if *Hymenoptera* should turn out as the sistergroup of the remaining *Endopterygota*, the *Diptera* species diverged around 150-100 Myr ago (personal communication with Rolf G.; (Yeates and Wiegmann, 1999; Wiegmann and Yeates, 2005; Beutel and Pohl, 2006)). The genomes of these species reveal considerable similarities (Kaufman *et al.*, 2002; Zdobnov *et al.*, 2002) but numerous differences also can be observed. Comparative analyses of the genomes of *Anopheles*, *Drosophila*, and *Apis* will be valuable for identifying for example bee genes that are lacking in the two dipterian genomes, some of which may be of importance for understanding bee specific features. Comparably, Anopheles has the ability to feed on the blood of the specific hosts. Hematophagy is essential for the female mosquito to produce eggs and propagate; it is also has been exploited by viruses and parasites that use Anopheles as a vehicle for transmission among vertebrates. Hematophagy is linked to specific host-seeking abilities as well as to nutritional challenges and requirements distinct from those of *Drosophila*.

Comparative genome studies of fruitfly and mosquito have been performed (Zdobnov *et al.*, 2002). In this work all the protein clusters were divided in groups: 1) ”one-to-one” orthologs, 2) ”many-to-many” orthologs, 3) homologues without easily discernable orthologous relationships and 4) proteins with no detectable homologs in any other species. On base of these data loss and the gain of genes were analyzed as well as expansions of protein families Our work is concentrated on the ”many-to-many” group of orthologs, and analyzing in-paralogs of every genome pairwise comparison, or genes family expansions. Furthermore, Zdobnov et al. used Clusters of Orthologous Groups (COG) approach (Tatusov *et al.*, 1997, 2003) for ortholog detection, whereas I used the Inparanoid software (Remm *et al.*, 2001; O’Brien *et al.*, 2005). Although both methods are able to find paralogs, the COG approach does not distinguish between in-paralogs and out-paralogs. Also COGs consist of at least three species. The Inparanoid approach is limited to two species and allows to define the evolutionary point of divergence and thus to separate out-paralogs from in-paralogs. Focusing on in-paralogs let me identify the duplicated genes and reveals adaptation processes on gene level which is characteristic for this species.

Species	Clusters of Orthologs	Number of In-paralogs for Species 1	Number of In-paralogs for Species 2
D. melanogaster A. gambiae	7414	11963	8890
D. melanogaster A. mellifera	5255	9290	9023
A. gambiae A. mellifera	5197	6637	9003

Table 4.1: Protein in-paralogs clusters of pairwise comparison between *Drosophila melanogaster*, *Anopheles gambiae* and *Apis mellifera*.
The clusters of orthologs with a single gene counterpart from each species ("one-to-one" orthologs) are representing 91-92% of the total number of clusters found by Inparanoid (tab. 4.2).

4.2 Results and Discussion

4.2.1 Duplication frequencies in Diptera and Hymenoptera

In-paralog detection was based on Inparanoid algorithm (Remm *et al.*, 2001; O'Brien *et al.*, 2005). As the calculation was performed on the proteome data, including splice variants of a gene, each protein was mapped onto its gene. As a result, groups of in-paralog genes were obtained. As a consequence of the mapping process, some clusters contained only a single gene for each species. Others, which were of importance for our analysis, contained one gene in the first and multiple genes in the second species, indicating a species specific expansion or loss. Additionally clusters with multiple genes from each species, based on independent duplication in each species, were found.

As fruitfly and mosquito are more closely related species the number of clusters of orthologs between them was found to be larger than when compared to bee. As the bee is equally distanced from fruitfly and mosquito on the phylogenetic tree the number of in-paralog proteins between bee and mosquito and bee and fruitfly is approximately the same (9003 and 9023, respectively) (tab. 4.1).

These clusters represent the "core" of the last common ancestor of both species. Com-

	Dr - An			Dr - Ap			An - Ap	
Dr1An2	360	912	Dr1Ap2	51	113	An1Ap2	48	137
Dr2An2	216	234	Dr2Ap2	63	47	An2Ap2	146	74
Dr2An1	577	244	Dr2Ap1	905	359	An2Ap1	1005	385
Single gene Counterparts	6737			4827			4741	
Total	7890	8127		5846	5346		5940	5337
Gene Clusters	7414			5255			5197	

Table 4.2: Gene clusters of pairwise comparison between *Drosophila melanogaster*, *Anopheles gambiae* and Apis mellifera (represented by genes numbers).
1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae*

pared to the total genes amount in *Drosophila-Anopheles* comparison single gene clusters would represent about 50% (6737 genes). This is in accordance with previous reports (Zdobnov *et al.*, 2002). For the comparison of *Drosophila* and *Anopheles* to *Apis* this number drops to about 35% percent of the total gene number. Fruitfly and mosquito are more closely related species and of course would have more gene homologs between themselves than with bee. Clusters containing at least one species specific duplication can be subdivided into the following three groups:

1. Multiple genes in species one and single gene in species two.

2. Single gene in species one and multiple genes in species two.

3. Multiple genes in both species.

4.2.2 Speciation events and duplications

The speciation between *Hymenoptera* and *Diptera* has happened about 250-200 Myr ago, whereas splitting Drosophila from *Anopheles* occurred 150-100 Myr ago. This timeframe allowed me to compare the duplication frequency in *Anopheles* and *Drosophila* before speciation and after. Therefore, I listed the genes duplicated in *Drosophila* compared to

Apis and searched for these genes in orthologs groups between *Anopheles* and *Drosophila.* If the genes were encountered in groups with duplicated Drosophila genes then the duplications have happened in the last 150 Myr. Contrarily, if the genes were encountered in ortholog groups with a single *Drosophila* gene the genes duplicated between 250-150 Myr ago. Finally, if the gene was not found at all in *Drosophila-Anopheles* ortholog groups then orthologs of this gene were either lost in *Anopheles* or newly invented in the *Diptera.* The same procedure was performed for the *Anopheles* genome. We found, that in *Drosophila* 40% (233 genes) of duplications happened in the common ancestor and the remaining 60% (364 genes) after speciation. For *Anopheles* these numbers shift slightly to 30% (200 genes) and 70% (522 genes), respectively. A similar increase of duplications was reported previously and explained by complications in the assembly process(Zdobnov *et al.*, 2002). In general these percentages are proportional to the evolution time.

Surprisingly, the amount of duplicated genes in *Apis* is very low (table 4.2), although bees show many unique physiological features. One explanation could be that the evolution of novel physiological features is linked to a very high evolutionary rate on the genome level. This finally might result in extremely divergent proteins which are beyond detectability by standard homology detection algorithms. Considering the high sensitivity of BLAST, this seems to be rather unlikely. Another explanation would be that the genomes of Drosophila and Anopheles are more derived than the one of *Apis.* From a human centric view this might be unexpected as we assume that the social lifestyle of bees would distinguish them from *Drosophila* and *Anopheles* also on the genome level. But, it seems that the evolution of for example hematophagy in Anopheles is the result of stronger evolutionary adaptation.

	Number of Gene Clusters		Fisher's exact test (P-value)
	Dr1	Dr2	
An1	6737	244	$2.42 \cdot 10^{-26}$
An2	360	73	
	Dr1	Dr2	
Ap1	4827	359	$9.73 \cdot 10^{-7}$
Ap2	51	18	
	An1	An2	
Ap1	4741	385	$1.62 \cdot 10^{-9}$
Ap2	48	23	

Table 4.3: Analyzes of gene clusters of pairwise comparison between *Drosophila melanogaster*, *Anopheles gambiae* and *Apis mellifera* (represented by clusters numbers).
1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae*

4.2.3 Independence of gene duplication

Our data allowed me to test, whether there is a correlation between genes duplicated independently in different species. Therefore, I analyzed the number of gene clusters for each pairwise comparison (tab. 4.1). In every comparison Fisher's exact test found p-values well below 0.001 indicating a strong association between variables. Thus, a gene duplicated in species one has a higher probability to be independently duplicated in species two than one not duplicated in species one. This might indicate, that, based on a stable evolutionary core of genes, the same type of genes are used for evolutionary adaptation in different species. This might be corroborated by a stronger correlation between Drosophila and Anopheles ($p = 2.42 * 10^{-26}$) compared to Diptera - Apis comparisons ($p = 9.73 * 10^{-7}$ and $p = 1.62 * 10^{-9}$)

All clusters of each pairwise comparison are divided in 4 groups: 1 to 1 orthologs, 1 to many - for species one and then species two, and the last group - many to many orthologs. For each of the subsets of data then Fischer's exact test is applied.

4.2.4 GO Classification of expanded genes

Having identified genes duplicated in a species, the next question arising was, what the function of these genes could be. The Gene Ontology Consortium has annotated genes in several model organisms using a controlled vocabulary of terms and placed the terms on a directed, acyclic graph (DAG). The three organizing principles of GO are cellular component, biological process and molecular function (Ashburner *et al.*, 2000; Harris *et al.*, 2004).

A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions. For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.

The cellular component ontology describes locations, at the levels of subcellular structures and macromolecular complexes. Examples of cellular components include nuclear inner membrane, with the synonym inner envelope, and the ubiquitin ligase complex, with several subtypes of these complexes represented.

Generally, a gene product is located in or is a subcomponent of a particular cellular component. The cellular component ontology includes multi-subunit enzymes and other protein complexes, but not individual proteins or nucleic acids. Cellular component also does not include multicellular anatomical terms.

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. As I was mainly interested in the broad functional classification, I restricted the analysis on GOSlim terms. GO identifiers were mapped directly to the genes and not to proteins. In cases where a gene resulted in multiple identifiers, the function which was represented mostly was assigned to this gene. If there was an equal number of GO identifiers for 2 or more functions, all of them were assigned to the gene. To date, detailed GO classification exists only for *Drosophila* and about 70% of the found genes had a GO annotation (tab. 3.3). In the clusters with a single *Drosophila* gene and multiple genes from the other species (*Apis* or *Anopheles*) these ortholog groups were characterized through the *Drosophila* genes. The clusters of orthologs were classified according to the cellular component, cellular process in which they are involved and molecular function (tab. 4.4). Caused by the annotation constraint, the most informative groups of clusters were those with multiple *Drosophila* and single

Anopheles or *Apis* genes. These cases might reveal, how the fruitfly evolved, which new proteins and function it acquired in comparison to bee and mosquito, or what was of importance for fruitfly survival and its unique appearance. Similarly, looking at the group of singular Drosophila gene and multiple Apis or Anopheles genes might indicate how bee or mosquito evolved.

Biological Process		Cellular Component		Molecular Function	
Biological Pr. Unknown	GO:0000004	Extracellular	GO:0005576	Nucleic Acid Binding	GO:0003676
Electron Transport	GO:0006118	Extracellular Matrix	GO:0005578	Motor Activity	GO:0003774
Nucleotide and Nucleic Acid Metabolism	GO:0006139	Extracellular Space	GO:0005615	Catalytic Activity	GO:0003824
Amino Acid and Derivative Metabolism	GO:0006519	Intracellular	GO:0005622	Helicase Activity	GO:0004386
Transport	GO:0006810	Cell	GO:0005623	Signal Transducer Activity	GO:0004871
Cell Motility	GO:0006928	Nucleus	GO:0005634	Receptor Activity	GO:0004872
Membrane Fusion	GO:0006944	Chromosome	GO:0005694	Structural Molecule Activity	GO:0005198
Cell Communication	GO:0007154	Cytoplasm	GO:0005737	Transporter Activity	GO:0005215
Development	GO:0007275	Unlocalized	GO:0005941	Carrier Activity	GO:0005386
Physiological Process	GO:0007582	Cellular Component Unknown	GO:0008372	Binding	GO:0005488
Behaviour	GO:0007610	Cell Surface	GO:0009986	Electron Transporter Activity	GO:0005489
Cell Growth and (or) Maintenance	GO:0008151	Membrane	GO:0016020	Protein binding	GO:0005515

Continue Table 4.4

Biological Process		Cellular Component		Molecular Function	
Metabolism	GO:0008152	External Encapsulating Structure	GO:0030312	Molecular Function Unknown	GO:0005554
Cell Death	GO:0008219			Aromatase Activity	GO:0008402
Catabolism	GO:0009056			Protein Transporter Activity	GO:0008565 (GO:0015463)
Biosynthesis	GO:0009058			Integrase Activity	GO:0008907
Pathogenesis	GO:0009405			Ion Transporter Activity	GO:0015075
Cellular Process	GO:0009987			Channel or Pore Class Transporter Activity	GO:0015267
Cell Differentiation	GO:0030154			Permease Activity	GO:0015646
Extracellular Structure, Organization and Biogenesis	GO:0043062			Antioxidant Activity	GO:0016209
Macromolecule Metabolism	GO:0043170			Kinase Activity	GO:0016301
Secretion	GO:0046903			Oxidoreductase Activity	GO:0016491
Regulation of Biological Pr.	GO:0050789			Transferase Activity	GO:0016740
Cellular Physiological Process	GO:0050875			Hydrolase Activity	GO:0016787
Response to Stimulus	GO:0050896			Lyase Activity	GO:0016829
				Isomerase Activity	GO:0016853
				Ligase Activity	GO:0016874
				Chaperone Regulator Activity	GO:0030188

Continue Table 4.4

Biological Process		Cellular Component		Molecular Function	
				Enzyme Regulator Activity	GO:0030234
				Transcription Regulator Activity	GO:0030528
				Translation Regulator Activity	GO:0045182

Table 4.4: GO Classification.

To test, whether there was a bias in the function of the duplicated genes, a chi-squared test of association (Pearson's chi-square test) was applied. This test is used to determine whether or not two variables measured on nominal or categorical scales are associated with each other and to determine whether a set of observed frequencies deviates significantly from a random model. Adjusted residuals describe both the strength and direction of this deviation and were used to identify functional classes strongly under- or overrepresented (tab. 4.5, 4.6 and 4.7). For all considered functional groups, the p-value was less then 0.001, indicating a significant dependency of accepted duplications to functional classes (see supplementary material for the complete table of genes for each of the GO identifiers and details of chi-square test results). Following, I detail out some functional groups highly correlated with duplicated genes.

4.2.4.1 Biological process

The largest groups of duplicated genes belonged in all species to the Physiological process (GO:0007582), Cellular process (GO:000987), Cellular physiological process (GO:0050875), Metabolism (GO:0008152) and Macromolecule metabolism (GO:0009987) categories (fig. 4.1). In Apis (set of genes Dr1Ap2) duplicated genes are involved in Cell communica-

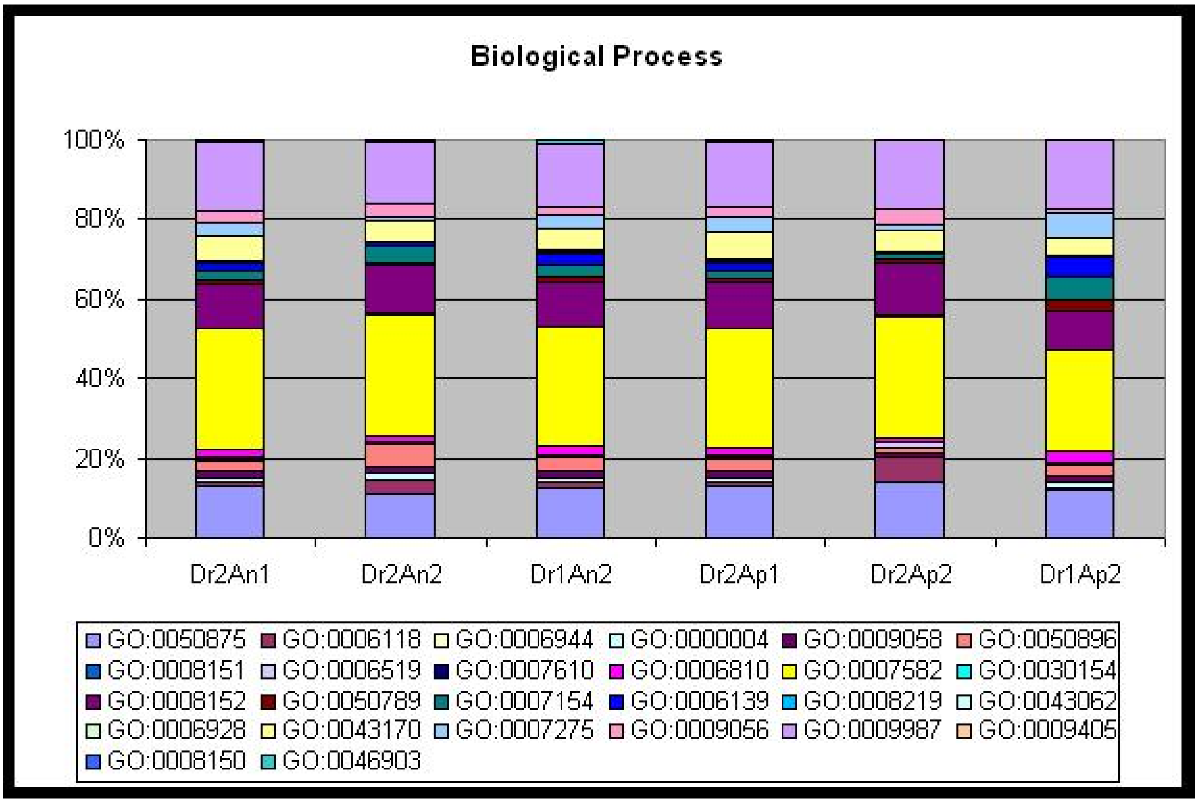


Figure 4.1: Percentage values of number of genes of different biological processes. 1 - corresponds to single gene; 2 - multiple genes; Dr - Drosophila melanogaster; Ap - Apis mellifera; An - Anopheles gambiae.

tion and Nucleotide and nucleic acid metabolism. Generally, Cell communicating genes, as well as Electron transport genes, were under evolutionary pressure for all of the 3 species. The Electron transport genes play an important role in metabolizing different pathogens and all of three compared species seem to have modified those mechanisms in the past 350 Myr.

4.2.4.2 Cellular location

The largest groups of genes are represented in the Cell category (GO:0005623) followed by Membrane (GO:0016020) and Intracellular (GO:0005622) categories (fig. 4.2). Analyzing the residuals (tab. 4.6) indicates that in *Drosophila* compared to *Anopheles* as well as vice versa duplicated genes' proteins preferentially belong to the extracellular

	Cellular Physiological Process	Electron transport	Biological Process Uknown	Biosynthesis	Response to Stimulus	Amino Acid Metabolism	Behaviour	Transport	Physiological Process	Cell Differentiation	Metabolism
Dr2An1	0,22	-1,50	0,51	-0,84	-0,68	-1,07	0,26	0,43	0,17	-1,12	-0,31
Dr2An2	-1,40	4,03	1,99	-0,16	4,04	0,50	-1,16	-1,53	0,12	0,75	0,70
Dr1An2	-0,14	-0,60	-0,78	0,08	0,95	-1,30	1,33	0,40	-0,03	0,30	-0,71
Dr2Ap1	0,42	-2,15	-0,42	1,18	-1,39	0,65	-0,11	0,41	0,03	-0,16	0,17
Dr2Ap2	0,75	6,71	-1,78	-1,52	-1,67	2,28	-0,72	-1,28	0,15	1,90	0,98
Dr1Ap2	-0,39	-0,83	0,37	-0,44	-0,18	0,09	-0,50	0,57	-1,22	-0,30	-0,66

	Regulation of Biol. Pr.	Cell Communication	Nucleotide and Nucleic Acid Met.	Cell Death	Extracellular Structure	Cell Motility	Macromolecule Metabolism	Development	Catabolism	Cellular Process
Dr2An1	0,14	-0,65	0,31	-0,80	-0,50	-0,12	0,27	0,11	0,55	0,75
Dr2An2	-2,51	3,18	-1,76	-1,78	-0,31	-1,42	-1,12	-3,73	1,59	-0,87
Dr1An2	1,79	0,74	1,44	0,94	-0,39	1,06	-1,04	0,42	-1,40	-0,65
Dr2Ap1	-0,19	-1,54	-0,23	1,09	0,82	0,51	1,33	1,48	-0,51	0,15
Dr2Ap2	-0,64	-1,03	-1,67	-1,10	-0,19	-0,88	-0,78	-1,58	1,58	0,20
Dr1Ap2	2,00	2,12	2,31	0,53	-0,13	-0,61	-0,98	1,88	-1,41	0,07

Table 4.5: Residual values for number of genes of different biological processes. Residual values more then +2 and less then -2 are marked with red. 1 - corresponds to single gene; 2 - multiple genes; Dr - Drosophila melanogaster; Ap - Apis mellifera; An - Anopheles gambiae.

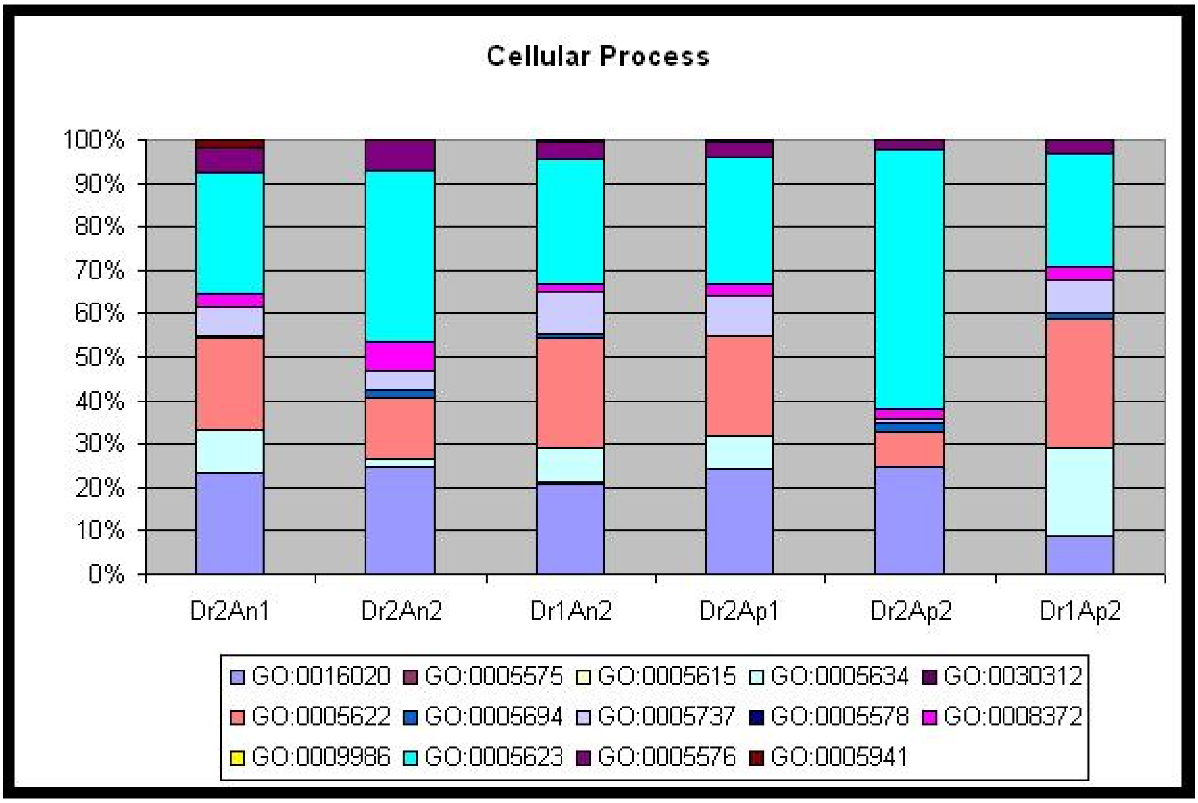


Figure 4.2: Percentage values of number of genes of different cellular locations. 1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae.*

space. *Anopheles*, for example, seemed to have gained within the process of evolving the ability to feed on blood a variety of mainly extracellular genes, which prevent platelet and clotting functions and modify inflammatory and immunological reactions in the vertebrate host. *Apis* shows underrepresentation of duplications of genes encoding Membrane proteins and active duplication of genes encoding Nucleus proteins.

4.2.4.3 Molecular function

For this branch of the Ontology, most duplicated genes belong to the Catalytic activity group (GO:0003824), followed by Binding (GO:0005488), Transporter activity (GO:0005215) and Hydrolase activity (GO:0016787) groups (fig. 4.3, tab. 4.7). High rates of genes duplications are observed among the genes of Antioxidant activity group.

	Membrane	Extracellular space	Nucleus	Intracellular	Chromosome	Cytoplasm
Dr2An1	0,38	-0,72	1,98	-0,27	-0,28	-1,37
Dr2An2	-0,27	-0,41	-3,70	-3,21	1,58	-2,15
Dr1An2	-1,24	3,06	0,31	1,49	0,89	1,31
Dr2Ap1	1,02	-0,95	-0,23	1,15	-1,59	1,79
Dr2Ap2	0,22	-0,24	-2,66	-2,87	1,74	-2,27
Dr1Ap2	-2,46	-0,21	3,77	1,33	0,77	-0,13
	Extracellular Matrix	Cellular Component Unknown	Cell	Extracellular	Unlocalized	
Dr2An1	1,40	0,15	-1,19	2,07	3,05	
Dr2An2	-0,50	2,71	1,55	1,68	-1,51	
Dr1An2	-0,68	-1,29	-0,79	-0,36	-0,56	
Dr2Ap1	-0,31	-0,40	-0,42	-1,72	-0,94	
Dr2Ap2	-0,29	-0,47	4,90	-0,93	-0,87	
Dr1Ap2	-0,25	-0,06	-0,57	-0,52	-0,76	

Table 4.6: Residual values for number of genes of different Cellular processes.
Residual values more then +2 and less then -2 are marked with red. 1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae.*

In *Drosophila* fewer duplications then expected were observed in the group of Structural molecule activity (GO:0005198), contrasting the set of genes Dr2An2 (duplication in *Drosophila* and *Anopheles*). Also overrepresented in Dr2An2 group are genes responsible for Ligase activity, Signal transducter activity, Oxidoreductase activity and Receptor activity. Contrasting, other classes in the duplicated genes, namely Transcription regulator activity, Kinase activity, Nucleic acid binding, Carrier activity and Transferase activity, are strongly underrepresented.

Groups of overrepresented genes in *Anopheles* (set of genes Dr1An2) belonged to Helicase activity, Channel or pore class transporter activity and Enzyme regulator activity groups.

In *Apis* (set of genes Dr1Ap2) duplicated genes belong to following groups: Helicase activity, Antioxidant activity and Structural molecular activity. Also fast evolving genes for both *Drosophila* and *Apis* (set of genes Dr2Ap2) belong to Motor and Oxidoreductase activity.

Carrier activity genes seemed to be actively duplicating in the common ancestor of Drosophila and Anopheles ($250 * 10^6 - 150 * 10^6$ years ago) but after the speciation event duplication of genes ceased.

4.2.5 Detailed examples

To investigate the molecular function of the duplicated genes in more detail I integrated them in to the cellular network via the KEGG database (Kanehisa *et al.*, 2004, 2006, 2008).

4.2.5.1 KEGG database

KEGG, or Kyoto Encyclopedia of Genes and Genomes, is a database resource for understanding higher-order functions and utilities of the biological system, such as the cell

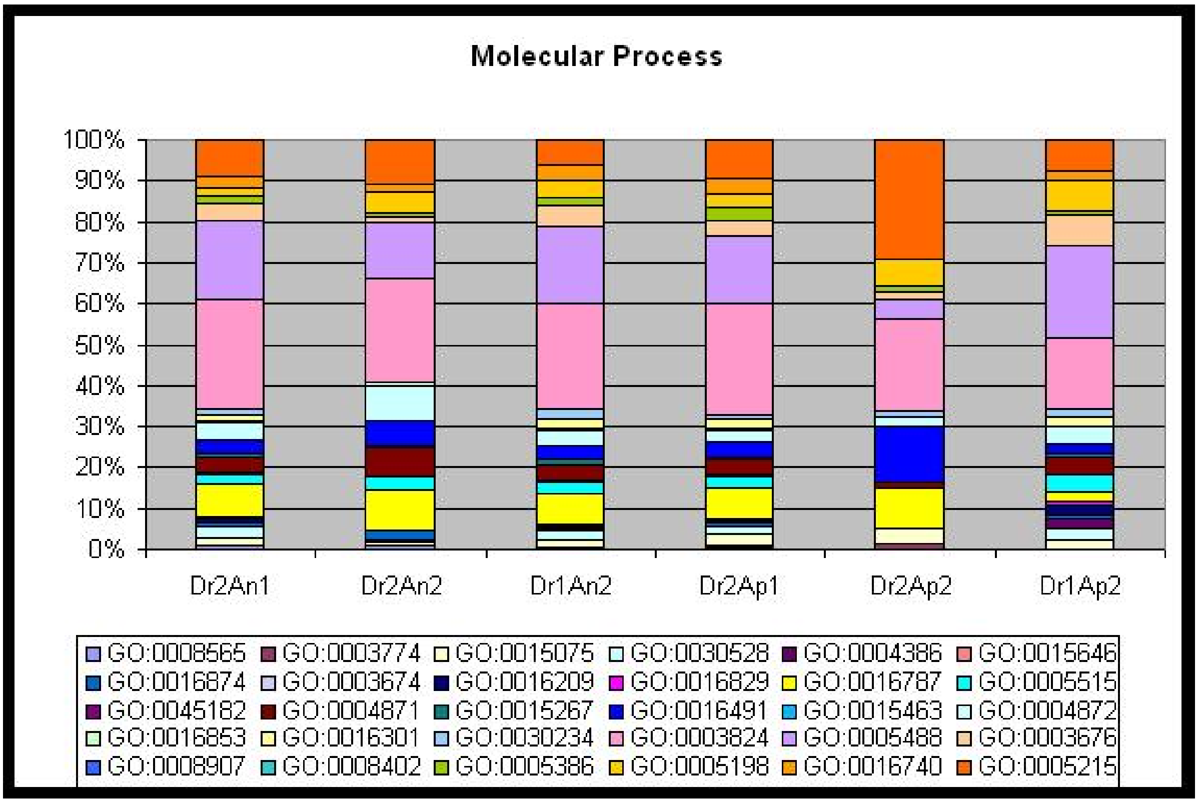


Figure 4.3: Percentage values of number of genes of different molecular processes. 1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae*.

	Protein Transporter Activity	Motor Activity	Ion Transporter Activity	Transcription Regulator Activity	Helicase Activity	Ligase Activity	Antioxidant Activity	Lyase Activity	Hydrolase Activity
Dr2An1	0,67	-0,82	-0,60	1,89	-1,23	-0,04	2,73	0,15	0,15
Dr2An2	0,34	-1,22	-1,96	-2,41	-0,80	3,41	-1,52	-0,66	1,64
Dr1An2	-0,33	-1,44	-1,01	0,67	3,31	-1,37	-1,80	-0,68	-0,16
Dr2Ap1	-0,07	1,51	1,71	-0,35	-1,64	-0,59	-0,58	0,72	-0,65
Dr2Ap2	-0,97	2,48	1,24	-1,75	-0,42	-1,14	-0,80	-0,94	0,95
Dr1Ap2	-0,77	-0,51	-0,20	0,77	5,65	0,20	2,49	0,59	-2,00

	Protein Binding	Translation Regulation Activity	Signal Transducer Activity	Channel or Pore Class Transporter Activity	Oxidoreductase Activity	Receptor Activity	Isomerase Activity	Kinase Activity
Dr2An1	-0,50	0,18	-0,90	1,21	-0,77	0,08	-0,84	-0,37
Dr2An2	1,42	-1,52	3,39	-0,50	2,13	5,59	0,92	-3,15
Dr1An2	-0,31	0,42	-0,16	2,51	-1,40	-0,10	-0,82	1,11
Dr2Ap1	0,19	0,70	-0,44	-2,03	-0,94	-2,53	1,01	1,56
Dr2Ap2	-2,01	-0,80	-1,72	-0,89	5,83	-0,84	-0,84	-1,66
Dr1Ap2	0,89	-0,64	-0,01	0,70	-1,01	0,08	-0,67	0,19

	Enzyme Regulator Activity	Catalytic Activity	Binding	Nucleic Acid Binding	Carrier Activity	Structural Molecule Activity	Transferase Activity	Transporter Activity
Dr2An1	0,67	0,25	1,71	1,23	-1,09	-3,41	-0,42	-0,81
Dr2An2	-1,93	-0,49	-1,75	-2,73	-2,32	2,62	-2,17	1,29
Dr1An2	2,51	-0,47	1,24	1,55	-1,18	1,54	0,29	-2,98
Dr2Ap1	-1,15	0,91	-0,44	-0,50	2,97	-0,55	1,92	-0,23
Dr2Ap2	0,00	-0,82	-3,70	-1,24	-0,73	2,11	-2,25	8,08
Dr1Ap2	0,65	-1,88	1,15	1,62	-0,77	2,04	-0,68	-0,70

Table 4.7: Residual values for number of genes of different Molecular processes.
Residual values more then +2 and less then -2 are marked with red. 1 - corresponds to single gene; 2 - multiple genes; Dr - *Drosophila melanogaster*; Ap - *Apis mellifera*; An - *Anopheles gambiae*.

or the organism, from genomic and molecular information. KEGG could be considered as a computer representation of the biological system, consisting of building blocks and wiring diagrams.

KEGG provides a reference knowledge base for linking genomes to life through the process of PATHWAY mapping, which is to map, for example, a genomic or transcriptomic content of genes to KEGG reference pathways to infer systemic behaviors of the cell or the organism.

KEGG consists of four main databases. They are categorized as building blocks in the genomic space (GENES databases) and the chemical space (LIGAND database), wiring diagrams in the network space (PATHWAY database) and ontologies for pathway reconstruction (BRITE database).

The KEGG GENES database is a collection of gene catalogs for all complete genomes and some partial, generated from publicly available resources. All genomes in KEGG GENES are subject to SSDB computation and given manual KEGG ortholog assignments. Each GENES entry contains cross-reference information to outside databases, including NCBI gi numbers, Entrez Gene IDs and UniProt accession numbers.

The KEGG PATHWAY database is a collection of manually drawn pathway maps for metabolism, genetic information processing, environmental information processing such as signal transduction, various other cellular processes and human diseases. All pathways are based on extensive survey of published literature.

Upon the gene entry as output I get a link to the pathways where this gene products are involved (fig. 4.4).

4.2.5.2 Metabolism

As revealed by the GO analysis, many duplicated genes belonged to the ”Electron transport” category. One of the ortholog clusters containing multiple Drosophila genes and single Apis or Anopheles contained genes CG3560 and CG17568. They are involved

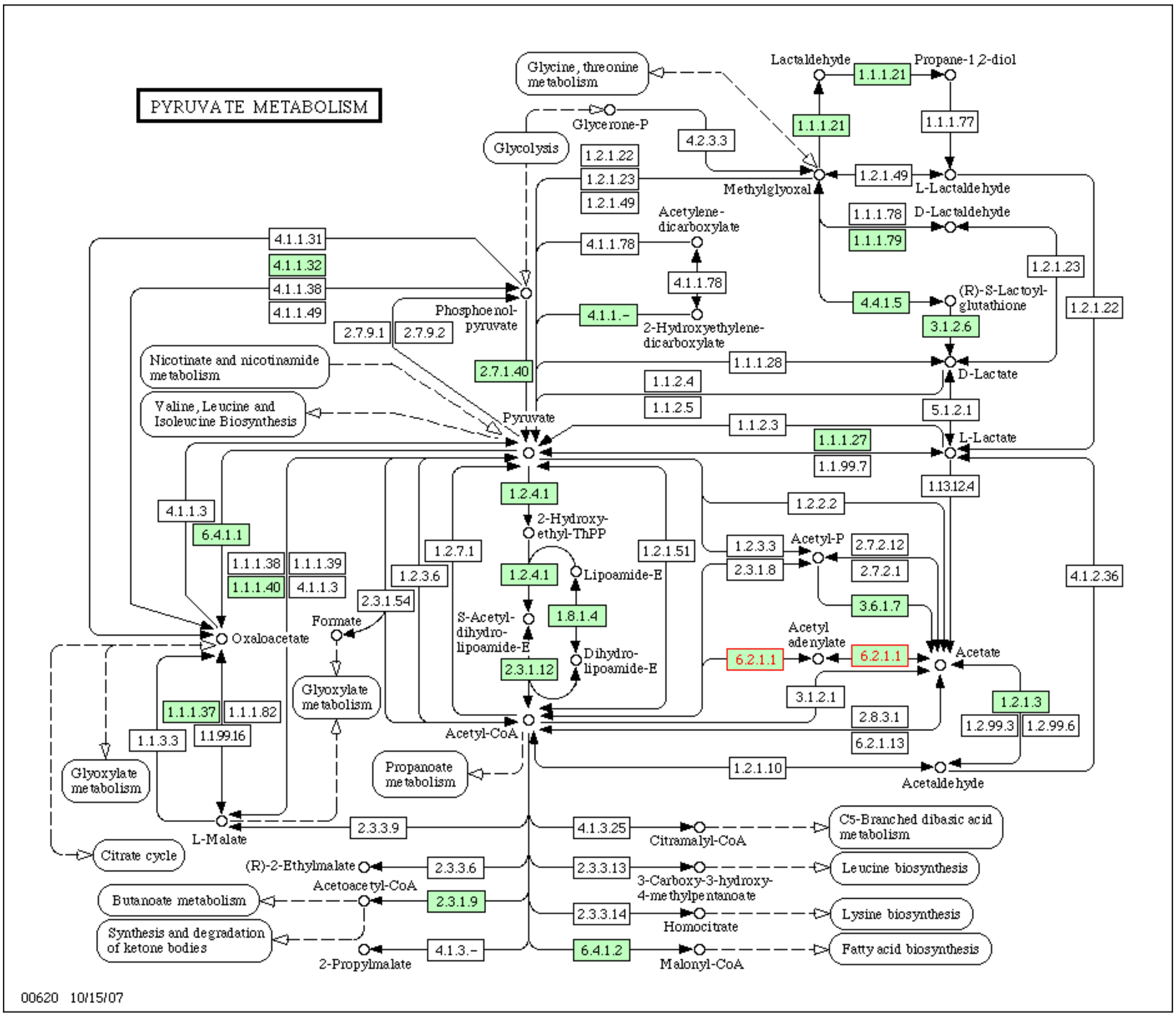


Figure 4.4: Pyruvate metabolism.
Product of Gene CG6432 acetyl-CoA synthetase (6.2.2.1) is involved in next pathways: dme00010 - Glycolysis/Gluconeogenesis, dme00620 - Pyruvate metabolism, dme00640 - Propanoate metabolism, dme00720 - Reductive carboxylate cycle (CO2 fixation).

in ubiquinol-cytochrome-c reductase activity and their products are part of oxydative phosphorylation cycle. Similarly, a lot of *Drosophila* in-paralogs corresponding to carbohydrates metabolism: (CG12055, CG32954, CG6432 - Glycolysis / Gluconeogenesis; CG4900 - Citrate Cycle (TCA); CG5103, CG8036 - Pentose phosphate pathway; CG8073, CG1982, CG10202, CG4649 - Fructose and Mannose metabolism; CG14934, CG14935, CG11669 - Galactose, Starch and Succrose metabolism). Duplication within these metabolic genes might reflect the sources of nutrients for *Drosophila* which are mainly the fruit juices and the yeast growing on rotting fruit.

4.2.5.3 Vision

The photoreceptors in *Drosophila* express a variety of rhodopsin isoforms (Harris et al. 1976; Stark et al. 2004). The R1-R6 photoreceptor cells express Rhodopsin1 (Rh1) which absorbs blue light (480 nm). The R7 and R8 cells express a combination of either Rh3 or Rh4 which absorb UV light (345 nm and 375 nm), and Rh5 or Rh6 which absorb blue (437 nm) and green (508 nm) light, respectively. Each rhodopsin molecule consists of an opsin protein covalently linked to a carotenoid chromophore, 11-cis-3-hydroxyretinal. The in-paralogs encoding these rhodopsin variety are CG10888 (opsin Rh3) and CG9668 (opsin Rh4). They were paralogues to the single *Anopheles* gene ENSANGG00000015219 and also to single *Apis* gene ENSAPMG00000007831. *Apis* gene encodes opsin which is ultraviolet sensitive with the maximum at the wavelength 350nm (Townson *et al.*, 1998; Spaethe and Briscoe, 2005). Thus, the duplication event led to the proteins specialized in more precise ultraviolet light absorption. Comparably, another ortholog group with multiple genes from *Drosophila* and *Apis* contains the Drosophila genes CG16740 (opsin Rh2) and CG4550 (opsin Rh1) and the *Apis* genes ENSAPMG00000000633 and ENSAPMG00000000632. As the latter encodes a green-sensitive opsin, this duplication might functions as an adaptation to the insects' need for vision during the day.

4.2.5.4 Scent

The ability to discriminate and respond to chemical signals from the environment is prerequisite for survival and plays extremely important role in the life cycles of *Apis*, *Drosophila* and *Anopheles*. It is known that odorant receptor genes have undergone massive duplication in *Anopheles* and *Drosophila* (Hill *et al.*, 2002). Accordingly, I found 4 ortholog groups containing multiple *Drosophila* and *Anopheles* genes encoding odorant receptors (or). The most massively expanded group contained 10 *Drosophila* and 17 *Anopheles* genes, encoding the whole range of odorant receptors. Such a relatively high number of duplication events in this gene category might indicate the importance of the odorant receptor genes in the last 150 Myr of evolution in these species. It was also shown that these genes have gone recent duplications in mammals (Emes *et al.*, 2004) where they are playing an important role in the process of feeding and mating habits.

4.2.5.5 Muscle structure

The duplications events can reflect adaptational process not only in adult but also in embryonal stage of development. As an example troponin genes in-paralogs in Drosophila are represented by CG7930 and CG9073. They are orthologs of ENSAPMG00000002676 in *Apis* and ENSANGP00000015945 in *Anopheles*. The *Drosophila* genes encode TpnC73F and TpnC47D, respectively. TpnC73F (TpnC Ia) shows a general, wide expression pattern, with a maximum level in abdominal hypodermal muscles and presents in embryonal and adult stage of development, whereas TpnC47D (TpnC Ib) is mainly expressed at the larval stage (Qiu *et al.*, 2003; Herranz *et al.*, 2005). This variation might allow fine-tuning of tissue-specific functions, and it has been demonstrated on a number of occasions that there is a functional non-equivalence between isoforms of structural muscle proteins (Fyrberg *et al.*, 1998).

4.2.6 Conclusion

In qualitative analyses I tried to establish the link between gene duplications and morphological features which are species speciefic. After grouping the in-paralogs and placing them on the timescale they were functionally grouped. For most of the in-paralogs, however, the function is still unknown.

I used two kinds of functional classifications: GO and KEGG database. There is a core of genes which are preferentially duplicated. The same genes are used in all three species. in other words only duplications in these genes are not rendered to become pseudogenes in the course of evolution.

Functional GO classification revealed over- and underrepresented groups. Some of the overrepresented groups could be linked with morphological features: electron transport gene are essential for the pathogen metabolizing. Overrepresentation of extracellular proteins in *Anopheles gambiae* is thought to reflect feeding on blood.

All in-paralogs were individually analyzed and mapped to KEGG pathway database. Individual genes involved in oxydative phosphorilation cycle, carbonhydrate metabolism, vision, scent and muscle structure were linked to morpholgical features of analysed insects.

5 Summary

In our analysis I was interested in the gene duplications, with focus on in-paralogs. In-paralogs are gene duplicates which arose after species split. Here I analysed the in-paralogs quantitatively, as well as qualitatively.

For quantitative analysis genomes of 21 species were taken. Most of them have vastly different lifestyles with maximum evolutionary distance between them 1100 million years. Species included mammals, fish, insects and worm, plus some other chordates. All the species were pairwised analysed by the Inparanoid software, and in-paralogs matrix were built representing number of in-paralogs in all vs. all manner.

Based on the in-paralogs matrix I tried to reconstruct the evolutionary tree using in-paralog numbers as evolutionary distance. If all 21 species were used the resulting tree was very far from real one: a lot of species were misplaced. However if the number was reduced to 12, all of the species were placed correctly with only difference being wrong insect and fish clusters switched. Then to in-paralogs matrix the neighbour-net algorithm was applied. The resulting "net" tree showed the species with fast or slow duplications rates compared to the others. We could identify species with very high or very low duplications frequencies and it correlates with known occurrences of the whole genome duplications.

As the next step I built the graphs for every single species showing the correlation between their in-paralogs number and evolutionary distance. As we have 21 species, graph for every species is built using 20 points. Coordinates of the points are set using the evolutionary distance to that particular species and in-paralogs number. In mammals with increasing the distance from speciation the in-paralogs number also increased, however not in linear fashion.

In fish and insects the graph close to zero is just the same in mammals' case. However, after reaching the evolutionary distances more than 800 million years the number of in-paralogs is beginning to decrease.

We also made a simulation of gene duplications for all 21 species and all the splits according to the fossil and molecular clock data from literature. In our simulation duplication frequency was minimal closer to the past and maximum in the near-present time. Resulting curves had the same shape the experimental data ones. In case of fish and insect for simulation the duplication rate coefficient even had to be set negative in order to repeat experimental curve shape.

To the duplication rate coefficient in our simulation contribute 2 criteria: gene duplications and gene losses. As gene duplication is stochastical process it should always be a constant. So the changing in the coefficient should be solely explained by the increasing gene loss of old genes. The processes are explained by the evolution model with high gene duplication and loss ratio.

The drop in number of in-paralogs is probably due to the BLAST algorithm. It is observed in comparing highly divergent species and BLAST cannot find the orthologs so precisely anymore.

In the second part of my work I concentrated more on the specific function of in-paralogs. Because such analysis is time-consuming it could be done on the limited number species. Here I used three insects: *Drosophila melanogaster* (fruitfly), *Anopheles gambiae* (mosquito) and *Apis mellifera* (honeybee).

After Inparnoid analyses and I listed the cluster of orthologs. Functional analyses of all listed genes were done using GO annotations and also KEGG PATHWAY database.

We found, that the gene duplication pattern is unique for each species and that this uniqueness is reflected through the differences in functional classes of duplicated genes. The preferences for some classes reflect the evolutionary trends of the last 350 million years and allow assumptions on the role of those genes duplications in the lifestyle of species. Furthermore, the observed gene duplications allowed me to find connections between genomic changes and their phenotypic manifestations. For example I found duplications within carbohydrate metabolism reflecting feed pattern adaptation, within photo- and olfactory-receptors indicating sensing adaptation and within troponin indicating adaptations in the development. Despite these species specific differences, O found high correlations between the independently duplicated genes between the species. This might hint for a "pool" of genes preferentially duplicated. Taken together, the observed duplication patterns reflect the adaptational process and provide us another link to the field of genomic zoology.

6 Zusammenfassung

In unserer Analyse untersuchten wir Genduplikationen mit besonderem Fokus auf "In-paralogen". In-paraloge sind Genduplikationen die nach Speziazion enstehen. Diese betrachteten wir hier in einer quantitativen als auch qualitativen Messreihe.

Die quantitative Analyse umfasste Genome aus insgesamt 21 Spezies. Der Großteil diese hat verschiedene Lebensgewonheiten mit eine maximalen Evolutionsdistanz von 1100 Millionen Jahren. Die Arten bestanden aus Säugetiere, Fischen, Insekten und Würmern, sowie weiteren Chordaten. Alle Arten wurden mittels der Inparanoid Software paarweise "all against all" analysiert und in in-paralog Matrizen gespeichert.

Basierend auf der in-paralog Matrix versuchten wir den evolutionären Baum über die Anzahl der In-paraloge als Maß für die evolutionäre Distanz zu rekonstruiren. Bei der Betrachtung alle 21 Arten würde der Baum jedoch sehr unpräzise: viel Arten wurden falsch plaziert. Durch eine Reduktion der Anzahl auf nur 12 Spezies clusterten jedoch alle Arten richtig, nur Insekten und Fische waren vertauscht. Anschließend wurde auf die In-paralog Matrix der Neighbor-net Algorithmus angewandt. Der daraus resultierende "Netz"-Baum repräsentiert die Spezies mit schneller oder langsamer Duplikationsrate im Vergleich zu den Anderen. Wir konnten Spezies mit sehr niedriger oder sehr hoher Rate identifizieren. Dabei korreliern die Genome mit der höheren Rate zu der Anzahl der auftauchenden Whole Genome Duplikationen.

Im nächsten Schritt erstellten wir Graphen für jede einzelne Spezies die das Verhältnis zwischen der Anzahl ihrer In-paraloger zur evolutionäre Distanz anzeigen. Jeder der

21 Graphen enthält insgesampt 20 Punkte. Die Punktkoordianten repräsentiern die evolutionere Distanz auf der X-Achse zu der Anzahl In-paraloger auf der Y-Achse. Bei Säugertieren wächst mit steigender Distanz auch die Anzahl In-paraloger. Das Verhältnis ist jedoch nicht linear.

Bei Fischen und Insekten ist der Graph in der Nähe des Nullpunkts gleich dem von Säugerteren. Beim Erreichen einer Distanz von mehr als 800 Millionen Jahren sinkt jedoch die Anzahl der In-paralogen.

Wir haben nun zusätzlich eine Simulation der Genduplikationen für alle 21 Spezies und alle dazu gehor̈igen Splits dürchgeführt. Die Splits wurden aus publizierten Fossilien und "Molecular Clock" Daten entnommen. In unsere Simulation stieg die Duplikationsrate mit Annäherung an die heutige Zeit. In Vergleich zu den Experementellen Daten haben die simulierten Graphen das gleiche Aussehen. Bei Fischen und Insekten musste der Koeffizient der Duplikationsrate negiert werden um die experimentelle Kurve zu erhalten. Der Koeffizient der Duplikationsrate stützt sich dabei auf folgende 2 Kriterien: Gen-Dupliaktion und Gen-Verlußt. Da Genduplikationen einem stochastischen Prozess folgen sollten sie immer konstant sein. Daher sind die erhöhten Genverlußte alter Gene verantwortlich für die Veränderunrg dieses Koeffizienten. Die Erklärung f”̆er dieses Verhalten basiert auf dem Evolutionsmodel - mit höhem Gen-Verlußt und hoher Gen Duplikation.

Der Verlußt der In-Paralogen enstehet wahrscheinlich durch den BLAST Algorithmus. Man beobachtet dies besonders bei sehr divergenten Arten bei dennen BLAST die Orthologen nicht mehr so prezise findet. Der zweite Teil meiner Arbeit bezieht sich auf die spezifische Funktion von In-paralogen. Da diese Analyse sehr zeitaufwendig ist konnte sie nur an einer begrenzten Anzahl von Spezies durchgeführt werden. Hier habe ich die folgenden drei Insekten verwendet: *Drosophila melanogaster* (Fruchtfliege), *Anopheles gambiae* (Moskito) und *Apis mellifera* (Honigbiene).

Alle durch die Inparanoid-Software entstandenen Cluster wurden mit der GO Annotation und der KEGG Pathway Datenbank analyiert.

Wir haben herausgefunden dass das Gen-Duplikationsmuster für jede Spezies einzigartig ist, und dass diese Einzigartigkeit durch Funktionale Unterschiede in duplizierten Genen entsteht. Die Bevorzugung einiger Gene repräsentiert die Evolutionsgeschichte der letzten 350 Millionen Jahre und erlaubt Annahmen über die Auswirkung der Gen Duplikationen im Leben der Spezies zu treffen. Weiterhin fanden wir durch die beobachteten Genduplikationen Zusammenhänge zwischen der Genomveränderung und ihrer phenotypischen Manifestation. Beispielsweise haben wir Duplikationen innerhalb des Karbohydratestoffwechsels für die Anpassung des Essvehaltens, Photo- und Olifaktorisch Rezeptoren - für Seh- und Geruchsvermögen und Troponin - zuständig für die Muskelentwicklung gefunden. Trotz diese speziesspezifischen Unterschiede haben wir starke Korrelation zwischen unabhängig duplizierten Genen erkannt. Dies könnte ein Indikator für einen "Pool" von bevorzugt duplizierten Genen sein. Zusammengefasst stellen die beobachteten Duplikationsmuster den Evolvierungsprozess dar, und liefern eine weitere Verbindung zur genomischen Zoologie.

7 References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, *et al.* (2000): The genome sequence of drosophila melanogaster. *Science*, **287** (5461): 2185–95.

Ainscough R, Bardill S, Barlow K, Basham V, Baynes C, Beard L, Beasley A, *et al.* (1998): Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, **282** (5396): 2012–8.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997): Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25** (17): 3389–402.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, *et al.* (2002): Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, **297** (5585): 1301–10.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, *et al.* (2000): Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25** (1): 25–9.

Ball CA and Cherry JM (2001): Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms. *Curr Opin Chem Biol*, **5** (1): 86–9.

Bandelt HJ and Dress AWM (1992): A canonical decomposition theory for metrics on a finite set. *Adv Math*, **92**: 47–105.

Beutel RG and Pohl H (2006): Head structures of males of strepsiptera (hexapoda) with emphasis on basal splitting events within the order. *J Morphol*, **267** (5): 536–54.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, *et al.* (2006): Ensembl 2006. *Nucleic Acids Res*, **34** (Database issue): D556–61.

Blair Hedges S and Kumar S (2003): Genomic clocks and evolutionary timescales. *Trends Genet*, **19** (4): 200–6.

Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, Noumen E, Pollet N, *et al.* (2008): Xenbase: a xenopus biology and genomics resource. *Nucleic Acids Res*, **36** (Database issue): D761–7.

Bryant D and Moulton V (2004): Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*, **21** (2): 255–65.

Cotton JA and Page RD (2005): Rates and patterns of gene duplication and loss in the human genome. *Proc Biol Sci*, **272** (1560): 277–83.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, *et al.* (2002): The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298** (5601): 2157–67.

Emes RD, Beatson SA, Ponting CP, and Goodstadt L (2004): Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Res*, **14** (4): 591–602.

Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, *et al.* (2008): Ensembl 2008. *Nucleic Acids Res*, **36** (Database issue): D707–14.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J (1999): Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151** (4): 1531–45.

Fyrberg EA, Fyrberg CC, Biggs JR, Saville D, Beall CJ, and Ketchum A (1998): Functional nonequivalence of drosophila actin isoforms. *Biochem Genet*, **36** (7-8): 271–87.

Gasterosteus aculeatus Fishbase (): ”gasterosteus aculeatus”. fishbase.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, *et al.* (2007): Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316** (5822): 222–34.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, *et al.* (2004): Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428** (6982): 493–521.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, *et al.* (2004): The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, **32** (Database issue): D258–61.

Harvey PH, May RM, and Nee S (1994): Phylogenies without fossiils. *Evolution*, **48** (3): 523–9.

He X and Zhang J (2005): Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169** (2): 1157–64.

Heger A and Ponting CP (2007): Evolutionary rate analyses of orthologs and paralogs from 12 drosophila genomes. *Genome Res*, **17** (12): 1837–49.

Herranz R, Mateos J, and Marco R (2005): Diversification and independent evolution of troponin c genes in insects. *J Mol Evol*, **60** (1): 31–44.

Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, *et al.* (2002): G protein-coupled receptors in anopheles gambiae. *Science*, **298** (5591): 176–8.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, *et al.* (2004): Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432** (7018): 695–716.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, *et al.* (2002): The genome sequence of the malaria mosquito anopheles gambiae. *Science*, **298** (5591): 129–49.

Huson DH and Bryant D (2006): Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, **23** (2): 254–67.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, *et al.* (2004): Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431** (7011): 946–57.

Jordan IK, Makarova KS, Spouge JL, Wolf YI, and Koonin EV (2001): Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res*, **11** (4): 555–65.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, *et al.* (2008): Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, **36** (Database issue): D480–4.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, *et al.* (2006): From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, **34** (Database issue): D354–7.

Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M (2004): The kegg resource for deciphering the genome. *Nucleic Acids Res*, **32** (Database issue): D277–80.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, *et al.* (2007): The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447** (7145): 714–9.

Kaufman TC, Severson DW, and Robinson GE (2002): The anopheles genome and comparative insect genomics. *Science*, **298** (5591): 97–8.

Kellis M, Birren BW, and Lander ES (2004): Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428** (6983): 617–24.

Kondrashov FA, Rogozin IB, Wolf YI, and Koonin EV (2002): Selection in the evolution of gene duplications. *Genome Biol*, **3** (2): RESEARCH0008.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, *et al.* (2004): Finishing the euchromatic sequence of the human genome. *Nature*, **431** (7011): 931–45.

Lespinet O, Wolf YI, Koonin EV, and Aravind L (2002): The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*, **12** (7): 1048–59.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, *et al.* (2005): Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438** (7069): 803–19.

Lynch M and Conery JS (2000): The evolutionary fate and consequences of duplicate genes. *Science*, **290** (5494): 1151–5.

Lynch M and Force A (2000): The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154** (1): 459–73.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S, Enard W, *et al.* (2005): Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437** (7055): 69–87.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, *et al.* (2007): Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature*, **447** (7141): 167–77.

Morin RD, Chang E, Petrescu A, Liao N, Griffith M, Chow W, Kirkpatrick R, *et al.* (2006): Sequencing and analysis of 10,967 full-length cdna clones from xenopus laevis and xenopus tropicalis reveals post-tetraploidization transcriptome remodeling. *Genome Res*, **16** (6): 796–803.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, *et al.* (2007): Genome sequence of aedes aegypti, a major arbovirus vector. *Science*, **316** (5832): 1718–23.

O'Brien KP, Remm M, and Sonnhammer EL (2005): Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, **33** (Database issue): D476–80.

Perriere G and Gouy M (1996): Www-query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78** (5): 364–9.

Qiu F, Lakey A, Agianian B, Hutchings A, Butcher GW, Labeit S, Leonard K, *et al.* (2003): Troponin c in different insect muscle types: identification of two isoforms in lethocerus, drosophila and anopheles that are specific to asynchronous flight muscle in the adult insect. *Biochem J*, **371** (Pt 3): 811–21.

R Development Core Team (2008): R: A language and environment for statistical computing.

Remm M, Storm CE, and Sonnhammer EL (2001): Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, **314** (5): 1041–52.

Small KS, Brudno M, Hill MM, and Sidow A (2007): A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome Biol*, **8** (3): R41.

Snelling WM, Chiu R, Schein JE, Hobbs M, Abbey CA, Adelson DL, Aerts J, *et al.* (2007): A physical map of the bovine genome. *Genome Biol*, **8** (8): R165.

Sonnhammer EL and Koonin EV (2002): Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, **18** (12): 619–20.

Spaethe J and Briscoe AD (2005): Molecular characterization and expression of the uv opsin in bumblebees: three ommatidial subtypes in the retina and a new photoreceptor organ in the lamina. *J Exp Biol*, **208** (Pt 12): 2347–61.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, *et al.* (2003): The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**: 41.

Tatusov RL, Koonin EV, and Lipman DJ (1997): A genomic perspective on protein families. *Science*, **278** (5338): 631–7.

Thomas D, Fowler C, and Hunt A (2005): *Programming Ruby: The Pragmatic Programmers' Guide*. Pragmatic Bookshelf.

Townson SM, Chang BS, Salcedo E, Chadwell LV, Pierce NE, and Britt SG (1998): Honeybee blue- and ultraviolet-sensitive opsins: cloning, heterologous expression in drosophila, and physiological characterization. *J Neurosci*, **18** (7): 2412–22.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala

R, *et al.* (2002): Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420** (6915): 520–62.

Weinstock GM, Robinson GE, A GR, Worley KC, Evans JD, Maleszka R, Robertson HM, *et al.* (2006): Insights into social insects from the genome of the honeybee apis mellifera. *Nature*, **443** (7114): 931–49.

Wiegmann BM and Yeates D (2005): *The Evolutionary Biology of Flies*. Columbia University Press, New York.

Wyder S, Kriventseva EV, Schroder R, Kadowaki T, and Zdobnov EM (2007): Quantification of ortholog losses in insects and vertebrates. *Genome Biol*, **8** (11): R242.

Yeates DK and Wiegmann BM (1999): Congruence and controversy: toward a higher-level phylogeny of diptera. *Annu Rev Entomol*, **44**: 397–428.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, *et al.* (2002): Comparative genome and proteome analysis of anopheles gambiae and drosophila melanogaster. *Science*, **298** (5591): 149–59.

Acknowledgements

First of all I want to thank prof. Jörg Schultz for supervising me during my whole time here in department of Bioinformatics. He guided me through all the problems and always was a grate motivator.

Prof. Thomas Dandekar for providing me with the opportunity to do my thesis in his department of Bioinformatics.

Dr. Tobias Müller and Dr. Roland Schwarz for helping with the statistics part.

Dr. Matthias Wolf for helping with phylogenetic trees.

Philipp Seibel for introducing me to the world of Ruby.

Frank Förster and Christian Koetschan for teaching me all about Latex.

All my roommates during my time in the Bioinformatics. They were the ones who helped me on the day to day basis with all my informatic problems.

All my colleagues in department of Bioinformatics for being my friends and creating a very cheerful atmosphere.

Curriculum vitae

I was born on the 28th of August 1980 in Minsk, Belarus From 1986 till 1994 I studied in the Secondary school #127 in Minsk, Belarus. From 1994 till 1996 - in Belarusian State University Lyceum. Starting in 1996 I was full-time student of Belarussian State Medical University and in 2002 successfully graduated with the major - general practioner. Since October 2002 till September 2004 I did M.D. training in the Laboratory of Biological Dosimetry, Clinic of Nuclear Medicine, University of Würzburg, Germany under the head of prof. Chr. Reiners. From 2004 till 2008 I was enrolled in the MD/PhD program of Würzburg University and working in the field of system biology. At current moment I am first year resident in the surgery department of Höchstadt hospital.

List of publications

Publications associated with this thesis

S. Vershenya, J. Schultz: In-paralogs Analysis of Insecta. Poster Presentation at *ISMB2006*, Fortaleza, Brasil, 6th-10th August, 2006

S. Vershenya, J. Schultz: Old Genes Die First. Poster Presentation at *ISMB2008*, Toronto, Canada, 20th-23th July, 2008

Prior Publications

S. Vershenya, J. Biko, V. Drozd, R. Lorenz, Chr. Reiners, K. Hempel: Dose Response For T-Cell Receptor (TCR) Mutants in Patients Repeatedly Treated with 131 I for Thyroid Cancer, *Mutation Research Regular Papers*, **548** (2004) pp 27-33

S. Vershenya, J. Biko, R. Lorenz, C. Reiners, H. Stopper, J. Grawe, K. Hempel: T-cell Receptor Assay and Reticulocyte-Micronuclei Assay as Biological Dosimeters for Ionizing Radiation in Humans, *Acta Med. Nagasaki*, **50** (2005) 15-21

H. Stopper, K. Hempel, Chr. Reiners, A. Heidland, S. Vershenya, R. Lorenz, V. Vukicevic, J. Grawe: Pilot Study For Comparison of Reticulocyte-Micronuclei with Lymphocyte-Micronuclei in Human Biomonitoring, *Toxicology Letters*, **156** (2005) pp 351-360

J. Grawe, J. Biko, R. Lorrenz, C. Reiners, H. Stopper, S. Vershenya, V. Vukecevic,

K. Hempel: Evaluation of the Reticulocyte Micronucleus Assay in Patients Treated with Radioiodine for Thyroid Cancer, *Mutation Research Regular Papers*, **583** (2005) pp 12-25