

# About the Interrelation of Evolutionary Rate and Protein Age

Hannes Luz

hannes.luz@molgen.mpg.de

Eike Staub

eike.staub@molgen.mpg.de

Martin Vingron

martin.vingron@molgen.mpg.de

Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

## Abstract

Evolutionary rate and gene age are interrelated when the age of a gene is assessed by the taxonomic distribution in the gene family. This is because homology detection by sequence comparison is depending on sequence similarity. We estimate family specific rates of protein evolution for orthologous families with representatives from man, fugu, fly, and worm. In fact, we observe that younger proteins tend to evolve faster than older ones. We estimate time points of duplication events that gave rise to novel protein functions and show that younger proteins were duplicated more recently than older ones.

**Keywords:** protein evolution, evolutionary rates, gene age, gene duplication

## 1 Introduction

Amino acid changing mutations in proteins are constrained by negative selection and accumulate at different rates. Rate variations among different proteins can be disentangled between different effects: rates vary among families, among lineages and among specific genes in specific lineages [26]. It is meanwhile accepted as a fact that rate variations are mainly due to family specific effects [14, 17]. For example, Jordan *et al.* [14] report that the rates of protein evolution are significantly correlated in different lineages of gene trees. That is, a fast or slow mutation rate is an attribute of the gene family. We confirmed that family specific evolutionary rates are meaningful quantities that can be assigned to individual protein families [17].

With a distribution of family specific evolutionary rates at hand, it is appealing to look into the principles that influence selection. Much research was done to establish relationships between gene indispensability, gene function and evolutionary rates [5, 7, 12, 17]. A study of Alba and Castresana [1] revealed the correlation of the age of a gene and its rate. The authors define the age of a gene by considering the taxonomic distribution of the genes in the family, that is, by the presence or absence of the gene in diverse lineages. This definition of gene age by its taxonomic distribution in the gene family is commonly used. For example, Kunin *et al.* [16] employ this definition and show that proteins of different ages obey distinct connectivity levels in interaction networks. Still, when protein families are inferred from levels of sequence similarity, we have to be aware of the fact that sequence similarity is expected to decrease exponentially in time. As a consequence, assigning an age by taxonomic distribution becomes intrinsically related to evolutionary rate. The latter was also demonstrated in a recent simulation study [9].

Another approach to assess the age of a protein is to trace back gene duplication events. The evolution of multigene families plays a fundamental role for the emergence of new gene functions and the time point of a duplication event can be regarded as the time where a new protein was invented during evolution [21].

In this study, we first confirm the result of Alba and Castresana [1] and show that proteins with a narrow taxonomic distribution are fast evolving. We then focus on multigene families and estimate time points of gene duplication events. We investigate whether there is a general relationship between the taxonomic range of proteins in multigene families and the time points of duplication events.

## 2 Age-Specific Rate Distributions

### 2.1 Family Specific Evolutionary Rates

We estimated family specific rates for 3640 sets of orthologs with representatives in the primate *Homo sapiens*, the pufferfish *Fugu rubripes*, the arthropode *Drosophila melanogaster*, and the nematode *Caenorhabditis elegans*. The peptide sequences were downloaded from the *Ensembl* database (version 16) [32]. We used the INPARANOID software [23] to derive sets of orthologs. We call these sets *orthologous families* and compute multiple alignments that contain a representative of each organism by DCA [28] for each orthologous family. We refer to the amino acid replacement model of Müller *et al.* [20] and to divergence time estimates from the literature [13, 30] in order to estimate a family specific rate  $\hat{\lambda}_X$  for an orthologous family  $X$  in a maximum likelihood framework. Units of family specific rates are PAM per Billions of Years (PAM/BYr). One PAM is the evolutionary distance where one substitution event per 100 sites is expected to have occurred according to the replacement model. Family specific rates range from 1 to 162 PAM/BYr. The mean rate amounts to 52 PAM/BYr, the median rate to 50 PAM/BYr. The derivation of orthologous families, multiple alignments family specific rates, and a set of extra-cellular families is described in detail in [17].

We observe that the rate distribution of 241 orthologous families with extra-cellular proteins is significantly shifted to larger rates [17]. Irrespective of this fact, extra-cellular proteins are supposed to have emerged relatively recently during the evolution of animals [6]. Disulfide bridges could not have formed in the earth's initial reducing atmosphere and abilities of proteins acting outside the cells must have gained significant impact during the evolution of multicellularity. Compounding both, the modernity of extra-cellular proteins and their elevated rates, gives rise to the hypothesis that the selective pressure acting on a protein is weaker the more modern or the younger the protein is. We call this hypothesis the *the-younger-the-faster-hypothesis*. In order to check whether the the-younger-the-faster-hypothesis holds in general, we aimed at assigning each orthologous family to an age class.

### 2.2 The Age of a Domain Architecture by its Taxonomic Distribution

While public sequencing has the number of new proteins grow exponentially, the number of domains found seems to be close to saturation [10]. There were likely no more than about 1000 protein domains in the primordial ancestors of present day organisms. Today's protein repertoire is assumed to have evolved from these domains [3, 4, 6]. Gene duplication followed by a recombination of protein domains is known to be a fundamental and fast process that continually gives rise to novel proteins. We argue that a protein that has a well defined function and is of a specific age obeys a specific domain architecture. Thus, we assign an age to an orthologous family by considering the taxonomic distribution of its domain architecture in the protein family.

Within the SYSTERS database we are given a partitioning of the publicly available proteins into disjoint clusters representing protein families [15, 18]. We mapped 3632 of 3640 orthologous families to a protein family from SYSTERS Release 4 under the requirement that the sequences within an orthologous family are present in the SYSTERS cluster. In SYSTERS clusters with multidomain proteins, it is sometimes observed that the proteins share a common domain but have a different domain architecture. For example, the domain architectures that are shown in Figure 4 of [17] that include the conserved Tyrosine Kinase domain all are present within one SYSTERS cluster. We subdivide a SYSTERS cluster into subclusters such that a subcluster contains sequences with the

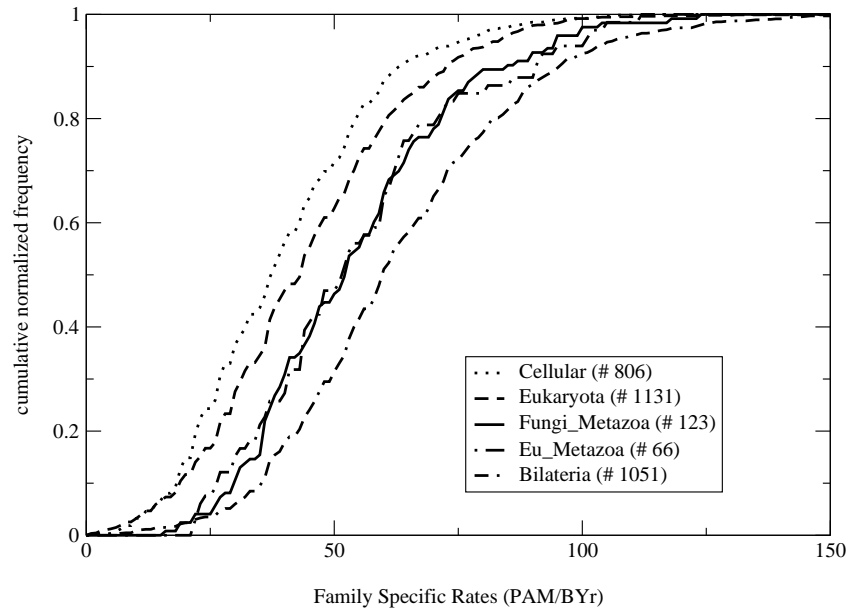


Figure 1: Normalized cumulative histograms of age-specific evolutionary rates.

same domain architecture.

We define a domain architecture as a sequence of domain occurrences where we count multiple consecutive occurrences of the same domain as one occurrence only. We searched the sequences of the orthologous families and the sequences of SYSTERS clusters for SMART and Pfam domains using “hmmpfam” [8], 7316 Pfam-HMMs from Release 12.0 [27], and 662 SMART-HMMs from Release 3.7 [25]. Predicted Pfam domains are accepted at a relatively weak E-value cutoff of  $E = 0.01$  (except for the Zn\_F domain that is accepted at the annotated E-value  $E = 10^{-4}$ ). SMART domains are accepted at the SMART provided E-value cutoffs for the lowest scoring true positive. If a detected SMART and Pfam domain covers the same region in a sequence, we prefer the SMART domain. Orthologous families are labeled with a domain architecture, if at least 3 of 4 sequences of the orthologous family have this domain architecture. We were able to label 3177 orthologous families with a domain architecture in this way.

The age class for an orthologous family corresponds to the least common internal taxon in the NCBI taxonomy when the sequences of the SYSTERS subcluster that have the domain architecture of the orthologous family are placed in the NCBI taxonomy [31]. In this way, orthologous families are assigned to either of the following age classes: *Cellular*, *Eukaryota*, *Fungi\_Metazoa*, *Eu\_Metazoa*, or *Bilateria*. For example, an orthologous family which is assigned to the age class *Cellular* has homologs with the same domain architecture in prokaryotes and the proteins are therefore supposed to be of ancient origin. Proteins of an orthologous family which is assigned to the age class *Eukaryota* are supposed to be “younger”. We call a set of orthologous families that belong to a specific age class an *age-specific* set. Since the sizes of the *Metazoa*- and the *Eumetazoa*-specific sets only amount to 30 and to 36 we merged these two sets into the composite *Eu\_Metazoa*-specific set.

### 2.3 Age-Specific Rate Distributions

We compared the rate distributions of age-specific sets. Figure 1 shows the normalized cumulative rate distributions and Table 1 shows mean rates and standard deviations for the age specific sets. The  $p$ -values in this table were computed by pairwise Wilcoxon two sample tests where the set in the row of the  $p$ -value was compared to the set one row above.

Table 1: Age-specific rate distributions.

age-specific subset	subset size	mean rate (PAM/BYr)	standard deviation (PAM/BYr)	<i>p</i> -value
Cellular	806	39.7	19.1	–
Eukaryota	1131	44.4	21.0	3e-7
Fungi_Metazoa	123	54.3	21.5	3e-6
Eu_Metazoa	66	54.3	22.5	0.919
Bilateria	1051	63.0	24.7	0.003

We observe that the mean rates obey the same order as the taxonomic units when traversing the taxonomy from *Cellular* to *Bilateria*. Except for the *Fungi\_Metazoa-Eu\_Metazoa* comparison the *p*-values of Wilcoxon two sample tests when subsequently comparing age-specific rate distributions are significant. It is likely that many families in the *Bilateria*-specific set were not classified as being *Metazoa*-specific because of the restricted number of sequences available for *Metazoa* which are not *Bilateria*. The latter might be a reason for the *p*-value of the *Fungi\_Metazoa-Eu\_Metazoa* comparison being close to 1.

Age-specific rate distributions support the the-younger-the-faster-hypothesis. Clearly, the rate distributions are subject to a systematic influence. We know that our ability to detect homology by sequence comparison is restricted by the amount of similarity the sequences obey. For example a fast evolving enzyme of ancient origin may have structural homologs in all kingdoms of life even if we are unable to detect remote homologies by sequence comparison. That is *Cellular*-specific sets are expected to include particularly slowly evolving proteins. Still, indirectly assessing homology relations by domain architectures is beneficial with respect to homology detection. Domains constitute the conserved parts of a protein and searching domains with Hidden Markov Models has been proven to be powerful and sensitive.

### 3 Multigene Families and Duplication Times

#### 3.1 Does “Age-specificity” Reflect Age?

The question remains open whether it is justified to say, for example, that the fast evolving proteins of the *Bilateria*-specific set have emerged more recently than the slowly evolving proteins in the *Cellular*-specific set. Our approach to address this question is an indirect one. We estimate time points of duplication events. If age-specificity reflects actual age, then the duplication events that gave rise to distinct orthologous families in our data set are supposed to have occurred earlier on average for the *Cellular*- and the *Eukaryota*-specific sets than for the *Metazoa*- and the *Bilateria*-specific sets.

Therefore, we analyzed multigene families where multiple copies of a protein with the same domain architecture are present in each of the four eukaryotic organisms under study. In other words, we consider multigene families where at least one duplication event occurred prior to the speciation events. The requirement that the orthologous families that make up such a multigene family have the same domain architecture is a practical one: we want to compute profile alignments of multiple alignments of orthologous families and measure evolutionary distances among paralogs. We can then extrapolate the evolutionary rates measured in the individual orthologous families and estimate duplication time points.

### 3.2 “Young” Genes were Duplicated more Recently than “Old” Genes

We define a multigene family as a set of orthologous families that obey the same domain architecture. and obtain 433 multigene families that are made up of 1689 orthologous families. The set of all multigene families is partitioned into an “old set”  $\mathcal{O}$  (*Cellular, Eukaryota*) and into a “young set”  $\mathcal{Y}$  (*Fungi\_Metazoa, Eu\_Metazoa, Bilateria*) according to the age-specificity of the orthologous families. While the size of the young  $\mathcal{Y}$  set amounts to 95, the old set  $\mathcal{O}$  comprises 338 multigene families.

For each pair of multiple alignments of orthologous families within a multigene family we compute a pairwise profile alignment using CLUSTALW [29] with default parameters. The evolutionary profile distance  $t_{XY}$  holds the average number of substitutions which have accumulated according to a replacement model between any two paralogous sequences between orthologous families  $X$  and  $Y$  [19]. Again, we refer to the Müller-Vingron amino acid replacement model [20] and estimate  $t_{XY}$  for each multigene family and for each pair of orthologous families in a multigene family.

Consider the ratio  $t_{XY}/2 \lambda_{XY}$  of the profile distance  $t_{XY}$  and the averaged family specific rate  $\lambda_{XY} = (\hat{\lambda}_X + \hat{\lambda}_Y)/2$  for orthologous families  $X$  and  $Y$ . This ratio is a rough estimate of the time of the duplication event that gave rise to the two orthologous families when we assume that  $\lambda_{XY}$  reflects the evolutionary rate of the sequences after the duplication event (see also Equation 1 and Figure 3 b). In order to check whether these ratios are different in  $\mathcal{Y}$  and  $\mathcal{O}$ , we compared both, distributions of evolutionary profile distances between orthologous families (that are paralogous) and distributions family specific rates in the two sets  $\mathcal{Y}$  and  $\mathcal{O}$ . We do this by having each multigene family contribute only once to the comparison: we assign to each multigene family  $i$  the arithmetic mean  $\bar{\lambda}_i$  of its family specific rates. Similarly, we average all pairwise profile distances within multigene family  $i$  and obtain  $\bar{t}_i$ .

Table 2: Evolutionary rates and evolutionary distances between paralogs in the set  $\mathcal{Y}$  and  $\mathcal{O}$ .

	#	average evolutionary family specific rates	average evolutionary paralog distance
young families $\mathcal{Y}$	95	60 PAM/BYr	250 PAM
old families $\mathcal{O}$	338	45 PAM/BYr	246 PAM
$p$ -value		$2.67 \cdot 10^{-10}$	0.717

In Table 2 we can see that the distribution of  $\bar{\lambda}_i$  in  $\mathcal{Y}$  significantly differs from the distribution of  $\bar{\lambda}_i$  in  $\mathcal{O}$ . The  $p$ -value of the Wilcoxon two sample test is  $2.67 \cdot 10^{-10}$ . While the average of  $\bar{\lambda}_i$ ,  $i \in \mathcal{Y}$  is  $\frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} \bar{\lambda}_i = 60$  PAM/BYr, the average of  $\bar{\lambda}_i$ ,  $i \in \mathcal{O}$  is  $\frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \bar{\lambda}_i = 45$  PAM/BYr. This result was expected and is in accord with the cumulative rate distributions of age-specific sets. Interestingly, the distributions of mean paralog distances do not significantly differ. The average of  $\bar{t}_i$ ,  $i \in \mathcal{Y}$  is  $\frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} \bar{t}_i = 250$  PAM and the average of  $\bar{t}_i$ ,  $i \in \mathcal{O}$  is  $\frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \bar{t}_i = 246$  PAM/BYr.

Figure 2 schematically compares phylogenetic trees for two representative multigene families, one of  $\mathcal{Y}$  and one of  $\mathcal{O}$ , that reflect the averaged values of Table 2. The trees are represented as dendrograms where the vertical axis scales with evolutionary distances or the number of amino acid replacement events respectively. Evolutionary distances to speciation events are larger for the “young” family. That is, the orthologs of the “young” family evolve at larger rates than the orthologs of the “old” family. On the other hand, the average evolutionary distance among paralogs is the same for the “old” and for the “young” family. Interpreting the situation with respect to a time scale, the duplication events within the “old” family occurred earlier than the duplication events for the “young” family. This result places an argument for the pertinence of assigning an age to a protein by the taxonomic distribution of the domain architecture in the protein family.

Are these observations due to the failure of detecting distant homologies for the families in  $\mathcal{Y}$  that evolve at large rates? The average number of orthologous families that are present in a multigene

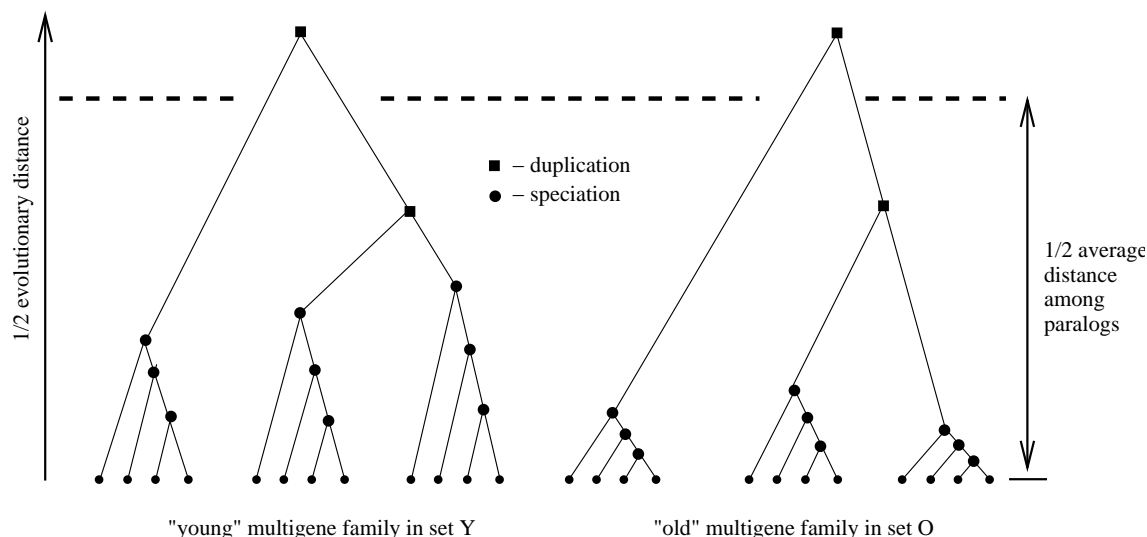


Figure 2: Schematic representation of multigene families. On the left the tree for a multigene family of the “young” set is shown. The tree on right represents a multigene family of the “old” set. Duplication events are marked by squares and speciation events are marked by circles. The vertical axis scales with evolutionary distances. The orthologs of the young family evolve at larger rates. The average evolutionary distance among paralogs is the same for the old and for the young set.

family of  $\mathcal{Y}$  amounts to 3.1 and to 4.1 in a multigene family of  $\mathcal{O}$ . We repeated the analysis and exclude the 8 largest multigene families with more than 20 orthologous families from  $\mathcal{O}$  such that the average number of orthologous families in  $\mathcal{O}$  decreases to 3.3. The results presented in Table 2 practically do not change. For instance, the average rate remains stable, the average paralog distance in  $\mathcal{O}$  becomes 245 PAM instead of 246 PAM and the  $p$ -value when comparing distance distributions is  $2.93 \cdot 10^{-10}$  instead of  $2.67 \cdot 10^{-10}$ .

In addition, we estimated concrete time points of early duplication events for pairs of orthologous families. The procedure requires a relative rate test where a third distantly related orthologous family as “outgroup” is involved. Note that there are no outgroup families for the most distantly related pairs of orthologous families. That is, when estimating duplication time points, we implicitly discard the most distant relationships.

### 3.3 Estimating Duplication Time Points

We pinpoint time points of duplication events that indicate the times where novel proteins have emerged. Consider two orthologous families  $X$  and  $Y$  that are present in the same multigene family. Any two sequences within  $X$  (or within  $Y$ ) per definition have diverged subsequent to speciation events. On the other hand, each sequence of  $X$  is related to each sequence of  $Y$  by a duplication event prior to the speciation events. When comparing paralogous sequences only, one has no clue of how to assess the time point of a duplication event. Yet, sometimes we can assume that sequences have evolved at constant rates. Then it is possible to extrapolate the rates measured on homologous regions between orthologs to infer the time of the duplication event. Rate constancy can be tested by a relative rate test [24]. The relative rate test requires that there is a third more distantly related outgroup family  $Z$  present in the multigene family (see Figure 3). In addition to the presence of an outgroup family  $Z$  in the multigene family, we also required that the age class of  $Z$  is the same or an older one than the age classes of  $X$  and  $Y$ .

We used CLUSTALW to compute the alignments that were necessary to perform the rate test and

to estimate duplication times. Profile alignments for pairs  $(X, Y)$  of orthologous families were already available (see Section 3.2). To obtain multiple alignments for three orthologous families  $(X, Y, Z)$ , we align the multiple alignment of the outgroup family  $Z$  to the profile alignment of  $X$  and  $Y$ . We call gapless subalignments in the profile alignment for  $(X, Y, Z)$  that comprise the four sequences of an orthologous family *orthologous profiles*  $x, y$  and  $z$  (see Figure 3a). The family specific rate measure [17] is applied to assess the rates  $\lambda_x$  and  $\lambda_y$  in orthologous profiles  $x$  and  $y$ . Under the assumption that  $\lambda_x$  and  $\lambda_y$  are constant in  $X$  and  $Y$ , we set  $\lambda_{xy} = (\lambda_x + \lambda_y)/2$ . The estimate of the duplication time  $\tau_{xy}$  is obtained from the relation

$$t_{xy} = 2 \cdot \lambda_{xy} \cdot \tau_{xy}. \quad (1)$$

The rate extrapolation is based on the assumption that the rate of protein evolution on homologous regions of the duplicated proteins remained approximately constant subsequent to the duplication as well as to the speciation events. The latter is checked by requiring the measured rates between orthologs to be close. To be precise we did not consider cases where  $\lambda_x$  and  $\lambda_y$  differed by a factor larger than 4/3. Rate constancy subsequent to the duplication event is tested by the relative rate test.

### The Relative Rate Test

For orthologous families  $X$  and  $Y$  in the presence of the more distantly related (outgroup) family  $Z$  we perform the relative rate test as follows. The relative rate test assumes that orthologous profiles  $x, y$  and  $z$  are the leaves in a phylogenetic tree with internal node  $i$  and that the inequalities  $t_{xz} > t_{xy}$  and  $t_{yz} > t_{xy}$  hold (see Figure 3a). Interpreting the three distances  $t_{xy}, t_{xz}$  and  $t_{yz}$  as path lengths and solving a system of three linear equations yields the three edge lengths  $t_{ix}, t_{iy}$  and  $t_{iz}$  of the tree. For example

$$t_{ix} = (t_{xy} + t_{xz} - t_{yz})/2.$$

If  $x$  and  $y$  have evolved from internal node  $i$  at a constant rate in a model tree, the edge lengths  $t_{ix}$  and  $t_{iy}$  are equal, i.e.,

$$\delta t := t_{ix} - t_{iy} = t_{xz} - t_{yz} = 0.$$

Since the accumulation of substitutions is subject to stochastic fluctuations and the measure of the profile distance comes with a measurement error,  $\hat{\delta t} = 0$  is almost never observed. Thus, we have to check whether the deviation of  $\hat{\delta t}$  from 0 is within the measurement error. To assess the uncertainties of the profile distance measures, 100 bootstrap replicates of the profile alignment of  $(X, Y, Z)$  were obtained and for each bootstrap replicate the respective profile distances were computed. First, we require that  $t_{xz} > t_{xy}$  and  $t_{yz} > t_{xy}$  are found for at least 95 bootstrap replicates. Second, we reason that, if the deviation of  $\hat{\delta t}$  from 0 is in the size of the measurement error, the probabilities that  $\hat{\delta t} < 0$  or that  $\hat{\delta t} \geq 0$  is observed should be approximately equal for one bootstrap replicate. A z-test is performed with the null-hypothesis  $H_0$  that the times where  $\hat{\delta t} < 0$  and  $\hat{\delta t} \geq 0$  are sampled from a binomial distribution with  $p_0 = \Pr(\delta t < 0) = \Pr(\delta t \geq 0) = 0.5$ . The relative rate test is passed if  $H_0$  can be accepted at a significance level of 95%, that is, if the number of times where  $\hat{\delta t} < 0$  is larger than 40 and smaller than 60 for 100 bootstrap replicates.

### Duplication Time Points

Computation of duplication times is restricted to 204 of 433 multigene families containing more than two orthologous families. These 204 multigene families comprise 1238 orthologous families. For each pair of orthologous families within a multigene family a profile distance is computed and rate tests are performed. Finally, we end with 115 estimated duplication times. The plot in Figure 3 compares the duplication times to the profile rates. The profile rates  $\lambda_{xy}$  are close to the family specific rates  $\lambda_X$  and  $\lambda_Y$ . Since they were calibrated by using the same set of divergence times that were used to calibrate the family specific rates [17], we do not observe instances where the estimated duplication times

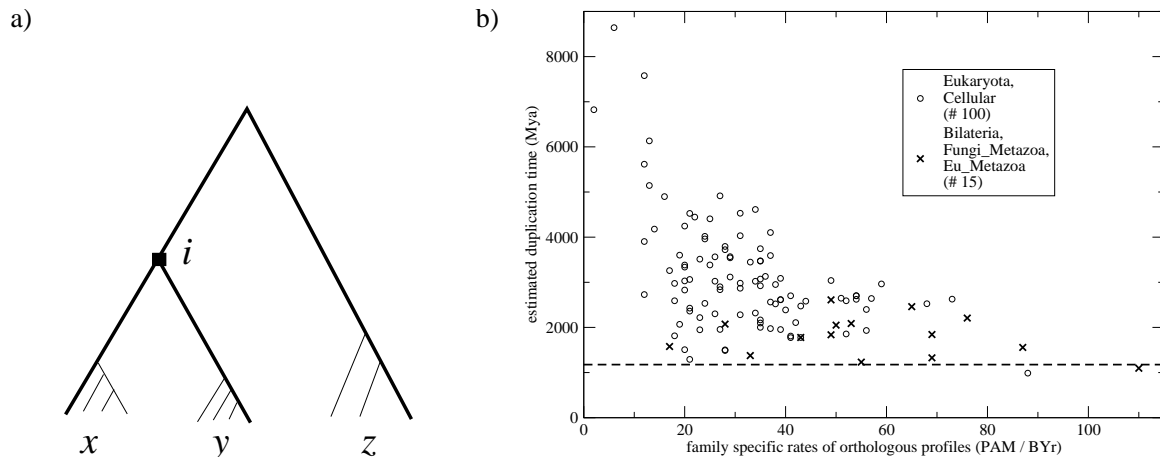


Figure 3: a) Relative rate test. Orthologous profiles  $x$ ,  $y$  and  $z$  are placed at the leaves of a phylogenetic tree. If sequences in  $x$  and  $y$  have evolved at a constant rate following the duplication, the profile distances  $t_{xz}$  and  $t_{yz}$  are equal. b) Estimated duplication time points compared to evolutionary rates measured on homologous regions of profile alignments between orthologous families. The dashed line is placed at 1170 Millions of years, the time of the nematode-arthropode split in the tree that we used to calibrate family specific rates [17].

are significantly smaller than 1170 Millions of years, the pre-given time for the assumed arthropode-nematode split [17]. Still, there are 18 estimated times exceeding 4 billions of years, the putative age of the earth. This might be caused by overestimated divergence times [11, 22] or a systematic decrease in evolutionary rates following the duplication event.

The scatter plot in Figure 3 is in accordance to the the-younger-the-faster-hypothesis. While duplication times of slowly evolving proteins cover a wide range, the upper bound of estimated duplication times decreases with increasing rates.

Further, we draw the connection of estimated duplication times to the age-specificity. Again, we divide the duplication times in an “old” set labeled with taxa *Cellular* and *Eukaryota* and a “young” set labeled with taxa *Fungi\_Metazoa*, *Eu\_Metazoa* and *Bilateria*. The “old” set contains 100 duplication times, the size of the young set is 15. Like in the analysis of all multigene families, the two distributions of profile distances which were used to assess the duplication times do not differ significantly. The pairwise Wilcoxon two sample test yields a  $p$ -value of  $p = 0.14$ . Yet the estimated duplication times for the old set are significantly larger than the duplication times for the young set ( $p = 1.80 \cdot 10^{-6}$ ).

The estimated duplication times support the view that genes in multigene families of a “young” age class were duplicated more recently than genes in multigene families of an “old” age class.

## 4 Summary and Discussion

We analyzed protein families comprising orthologous sequences from man, fugu, fly and worm. Family specific evolutionary rates were measured by applying standard maximum likelihood tree estimation procedures. Since modern extra-cellular proteins evolve at elevated rates, we are motivated to investigate the the-younger-the-faster-hypothesis. The hypothesis states that the evolutionary rate of a protein tends to be larger the more recently the protein emerged in evolution.

We assigned families of orthologs to age classes which reflect the taxonomic distribution of proteins with the same domain architecture. Age-specific rate distributions are in accordance with the hypothesis. However, because homology detection depends on evolutionary rates, we cannot decide,

whether the-younger-the-faster holds or whether we should rather say: the faster a protein evolves, the younger it seems to be.

Finally, we were able to place an argument for the pertinence of assigning an age to an orthologous family by the taxonomic distribution of domain architectures. We traced back duplication events which gave rise to the emergence of novel protein functions. It turns out that duplication times are relatively small for multigene families from “young” age classes and relatively large for multigene families from “old” age classes. The latter confirmed our assignment of the labels “young” and “old” to these classes. We found that evolution has been acting at a different rate since the split of man, fish, fly, and worm for orthologous families with labels “young” and “old”. These results support the the-younger-the-faster-hypothesis.

Some duplication time estimates predating the earth’s putative origin demand discussion. First, the overestimation of duplication times might be due to overestimated divergence times that were used to calibrate family specific rates. Indeed, time estimates for the emergence of animals that are obtained from molecular data vary by a factor of 2 and range from 550 to 1170 Millions of years [11, 22]. The divergence times that we used for calibration are at the upper limit. Second, the relative rate test is a necessary but not a sufficient requirement for rate constancy to hold. Duplication times are overestimated when evolutionary rates along different lineages simultaneously decrease in time.

Alba and Castresana provide two interpretations for the observed interrelation of evolutionary rate and gene age [1]. The first one states that the selective pressure is weaker for new genes. The second one assumes, that the tolerance of a gene against accepting mutations decreases in time which means that the selective pressure is stronger for old genes. This study is in favor of the latter interpretation. Because, if the evolutionary rate of a gene decreases in time, divergence times are systematically overestimated. Overestimated divergence times in turn would explain the overestimated duplication times.

## Acknowledgments

We want to thank to Thomas Meinel for his help with the NCBI taxonomy in SYSTERS, Jörg Schultz for his help regarding the computation of SMART E-value cutoffs, and Tobias Müller for valuable discussions about profile distances.

## References

- [1] Alba, M.M and Castresana, J., Inverse relationship between evolutionary rate and age of mammalian genes, *Mol. Biol. Evol.*, 22(3):598–606, 2005.
- [2] Bairoch, A. and Apweiler R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, 28:45–48, 2000.
- [3] Chothia, C., Protein families in the metazoan genome, *Dev. Suppl.*, 27–33, 1994.
- [4] Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A., Evolution of the protein repertoire, *Science*, 300(5626):1701–1703, 2003.
- [5] Castillo-Davis, C.I., Kondrashov, F.A., Hartl, D.L., and Kulathinal, R.J., The functional genomic distribution of protein divergence in two animal phyla: Coevolution, genomic conflict, and constraint, *Genome Res.*, 14(5):802–811, 2004.
- [6] Doolittle, R.F., The multiplicity of domains in proteins, *Annu. Rev. Biochem.*, 64:287–314, 1995.
- [7] Drummond, D.A., Raval, A., and Wilke C.O., A Single Determinant Dominates the Rate of Yeast Protein Evolution, *Mol. Biol. Evol.*, 23(2):327–337, 2006.

- [8] Eddy S.R., Profile hidden Markov models, *Bioinformatics*, 14(9):755–763, 1998.
- [9] Elhaik, E., Sabath, N., and Graur, D., The “inverse relationship between evolutionary rate and age of mammalian genes” is an artifact of increased genetic distance with rate of evolution and time of divergence, *Mol. Biol. Evol.*, 23(1):1–3, 2006.
- [10] Geer, L.Y., Domrachev, M., Lipman, D.J., and Bryant, S.H., CDART: Protein homology by domain architecture, *Genome Res.*, 12(10):1619–1623, 2002.
- [11] Graur, D. and Martin W., Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision, *Trends Genet.*, 20(2):80–86, 2004.
- [12] Hahn, M.W. and Kern, A.D., Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol. Biol. Evol.*, 22(4):803–805, 2005.
- [13] Hedges, S.B., The origin and evolution of model organisms, *Nature Review Genetics*, 3(11):838–849, 2002.
- [14] Jordan, I.K., Kondrashov, F.A., Rogozin, I.B., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V., Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins, *Genome Biol.*, 2(12):RESEARCH0053, 2001.
- [15] Krause, A., Stoye, J., and Vingron, M., Large scale hierarchical clustering of protein sequences, *BMC Bioinformatics*, 6(1):15, 2005.
- [16] Kunin, V., Pereira-Leal, J.B., and Ouzounis, C.A., Functional evolution of the yeast protein interaction network, *Mol. Biol. Evol.*, 21(7):1171–1176, 2004.
- [17] Luz, H. and Vingron, M., Family specific rates of protein evolution, *Bioinformatics*, 22(10):1166–1171, 2006.
- [18] Meinel, T., Krause, A., Luz, H., Vingron, M., and Staub, E., The SYSTERS protein family database in 2005, *Nucleic Acids Res.*, 33(Database issue):D226–D229, 2005.
- [19] Müller, T., Rahmann, S., Dandekar, T., and Wolf, M., Accurate and robust phylogeny estimation based on profile distances: A study of the Chlorophyceae (Chlorophyta), *BMC Evol. Biol.*, 4(1):20, 2004.
- [20] Müller, T., Spang, R., and Vingron, M., A comparison of dayhoff’s estimator, the resolvent approach and a maximum likelihood method, *Mol. Biol. Evol.*, 10(1):8–13, 2002.
- [21] Ohno, S., Ancient linkage groups and frozen accidents, *Nature*, 244(5414):259–262, 1973.
- [22] Peterson, K.J., Lyons, J.B., Nowak, K.S., Takacs, C.M., Wargo, M.J., and McPeck, M.A., Estimating metazoan divergence times with a molecular clock, *Proc. Natl. Acad. Sci. USA*, 101(17):6536–6541, 2004.
- [23] Remm, M., Storm, C.E., and Sonnhammer, E.L., Automatic clustering of orthologs and inparalogs from pairwise species comparisons, *J. Mol. Biol.*, 314(5):1041–1052, 2001.
- [24] Sarich, V.M. and Wilson, A.C., Generation time and genomic evolution in primates, *Science*, 179(78):1144–1147, 1973.
- [25] Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P., SMART, a simple modular architecture research tool: Identification of signaling domains, *Proc. Natl. Acad. Sci. USA*, 95(11):5857–5864, 1998.

- [26] Smith, N.G. and Eyre-Walker, A., Partitioning the variation in mammalian substitution rates, *Mol. Biol. Evol.*, 20(1):10–17, 2003.
- [27] Sonnhammer, E.L., Eddy, S.R., and Durbin, R., Pfam: A comprehensive database of protein domain families based on seed alignments, *Proteins*, 28(3):405–420, 1997.
- [28] Stoye, J., Moulton, V., and Dress, A.W., DCA: An efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment, *CABIOS*, 13(6):625–626, 1997.
- [29] Thompson, J.D., Higgins, D.G., and Gibson, T.J., CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22(22):4637–4680, 1994.
- [30] Wang, D.Y., Kumar, S., and Hedges, S.B., Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi, *Philos. Trans. R. Soc. Lond. Series B Biological sciences*, 266:162–171, 1999.
- [31] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., and Wagner, L., Database resources of the national center for biotechnology, *Nucleic Acids Res.*, 31(1):28–33, 2003.
- [32] <http://www.ensembl.org>