4 Protein evolution, the younger the faster

4.1 Metazoan radiation, the younger the faster?

We observe that extra-cellular proteins evolve at elevated rates. Irrespective of this fact, extra-cellular proteins are supposed to have emerged relatively recently during the *metazoan radiation* [Doolittle, 1995]. 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, metazoan radiation and the elevated rates of extra-cellular proteins, it stands to reason if there is a general relationship between the selective pressure acting on a protein and the time that has passed since its evolutionary emergence.

The experiments presented in this chapter were performed to investigate interrelations of the metazoan radiation, the emergence of proteins and evolutionary rates. In Section 4.2 we investigate the set of extra-cellular proteins in more detail. A case study focuses on the rates of protein tyrosine kinases and tyrosine kinase receptors. For the latter it is revealed that extra-cellular domains compared to their cytoplasmic counterparts are more divergent. This suggests that the range of accepted mutations in a protein's amino acid sequence is larger the more modern or the younger the protein is. We set up a working hypothesis and call it shortened "the younger the faster".

Addressing the hypothesis requires the assignment of an age to a protein. This is naturally done by considering the taxonomic distribution of sequences within the respective protein family [Meinel *et al.*, 2003; Kunin and Ouzounis, 2003]. For example, Kunin *et al.* considered the taxonomic range of protein families to show that proteins emerging at certain evolutionary periods obey distinct connectivity levels in interaction networks [Kunin *et al.*, 2004]. Yet, if 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 and assigning an age becomes intrinsically related to evolutionary rate.

The dependance of homology detection by sequence comparison on evolutionary rates is exposed when we infer the ancient origin of proteins by searching the orthologous families against prokaryotic organisms (Section 4.3). In Sections 4.4 and 4.5, we further investigate the hypothesis. First by projecting orthologous families to a least common taxon. Second by considering evolutionary distances among paralogs in multigene families and by estimating duplication time points.

4.2 Extra-cellular proteins

4.2.1 Inferring extra-cellular localization

We combine different *in silico* approaches to derive an extended set of extra-cellular families with high accuracy (or a reduced false positive rate). Namely we check putative extra-cellular localization due to Swiss-Prot annotations, the detection of extra-cellular SMART domains and the existence of a predicted signal peptide or a transmembrane helix. If the proteins of an orthologous family meet at least two of those criteria, we call the respective family *extra-cellular*. Note that the derived set of extra-cellular families also includes proteins that are only in part extra-cellular (transmembrane proteins) but follow the same secretory pathway.

- **Textmining SwissProt entries**

  Nair and Rost [2002a] developed a fully automated method that analyzes Swiss-Prot keywords and predicts sub-cellular localization. They provide a list of keywords with a strong correlation to the annotated protein being localized in a specific sub-cellular compartment [Nair and Rost, 2002b]. We refer to this list and mark a protein as putatively extra-cellular, if a keyword in the Swiss-Prot entry points to an extra-cellular locale of the protein. Further we take the following features of a Swiss-Prot entry into account: Annotations of disulfide bridges not followed by the word "similar", comment lines where the term "secreted" occurs and cross-references to the database of Gene Ontologies [Harris *et al.*, 2004] that assign the protein to the extra-cellular space.

- **Localization of SMART domains, domain projection**

  SMART domains are searched as described in Section 3.6.6. In addition, when marking a protein as putatively extra-cellular due to a detected SMART domain, the results of the domain projection method [Mott *et al.*, 2002] are taken into account.

- **Detection of signal peptides and transmembrane helices**

  We search the sequences for signal peptides using SignalP. Protein sequences predicted from draft genomes often lack N-terminal regions. The prediction of transmembrane helices using TMHMM complements the prediction of signal peptides.

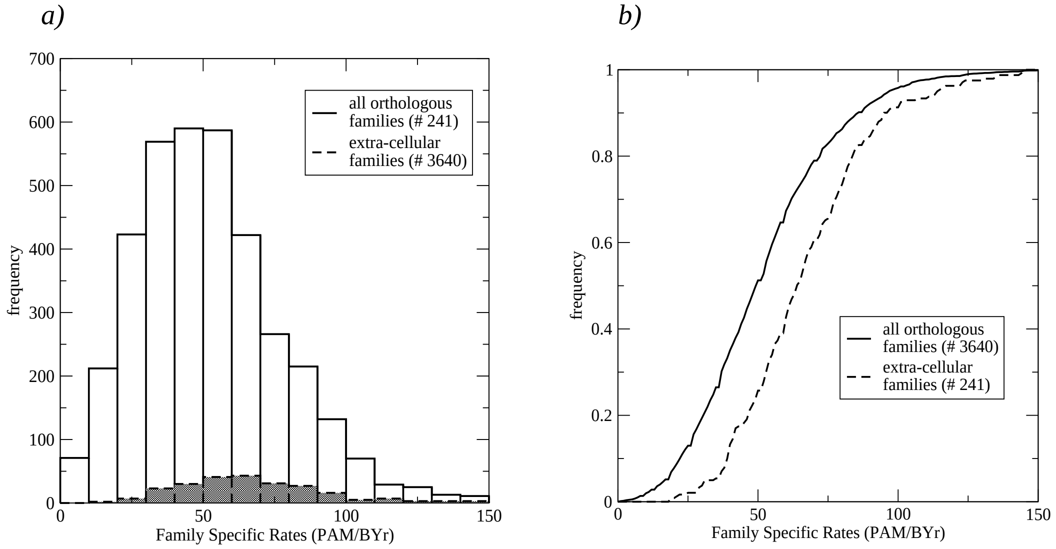


Figure 4.1: a) The distribution of Family Specific Rates for all orthologous families and for the set of extra-cellular families. b) The same distributions shown as cumulative normalized histograms.

4.2.2 Modern extra-cellular proteins are fast evolving

We end up with a set of 241 orthologous families with extra-cellular proteins. This set also includes transmembrane proteins which follow the secretory pathway but are only in part extra-cellular. As expected, we observe that the rate distribution of the extra-cellular families is significantly shifted to larger rates (see Figure 4.1). The mean rate and the median rate are 67.2 PAM/BYr and 64 PAM/BYr respectively (see Table 4.1). A Wilcoxon two sample test that compares the rate distribution of all orthologous families to the rate distribution of the extra-cellular families yields a significant p-value of $1.80 \cdot 10^{-19}$.

4.2.3 Evolution of protein tyrosine kinases (PTKs) and tyrosine kinase receptors (rPTKs)

Protein Tyrosine Kinases

Protein Tyrosine Kinases (PTKs) are involved in cellular signalling pathways and regulate key cell functions such as proliferation, cell growth, immune response and differentiation. Mutations in PTKs often play a significant role in diseases like diabetes and cancer. The three dimensional structure of the proto-oncogene Src is shown in Figure 2.3. The catalytic *Tyrosine Kinase* domain of PTKs phosphorylates specific tyrosine side chains.

We studied the evolution of the large PTK multigene family in greater detail and analyzed the 14 domain architectures shown in Figure 4.2. Each of the domain architectures is present within one SYSTERS cluster and in a distinct orthologous family.

Comparing PTKs is interesting with regard to a putative interrelation of a protein's extra-cellular localization and its evolutionary rate. While non-receptor PTKs are purely cytoplasmic, receptor PTKs (rPTKs) are membrane anchored and contain an extra-cellular ligand binding domain. Figure 4.2 shows the domain architectures of the PTKs. The set of 14 orthologous families divides into 9 families with receptor PTKs shown above the dashed line and 5 families containing non-receptor PTKs shown below the dashed line.

Evolutionary trees

Alignments of the Tyrosine Kinase domains are used to construct evolutionary trees of PTKs and to group the enzymes into subfamilies. For example, two such trees are provided in studies of Hanks and Hunter [1995] and Robinson *et al.* [2000].

We aligned the Tyrosine Kinase domains of 108 sequences in our data set using "hmmalign" and constructed bootstrapped NJ-trees as well as a Profile NJ (PNJ) tree. Domains of orthologous families cluster together in the NJ-trees. The bootstrap support at internal edges that partition groups of orthologous families is improved when the subalignments of orthologous families are used to construct the PNJ tree [Müller *et al.*, 2003, 2004]. Again, we choose the Müller-Vingron model as amino acid replacement model.

The PNJ-tree is shown on the left in Figure 4.2. Only edges with a bootstrap support being larger than 80% are resolved. The PNJ-tree is an unrooted tree. Yet, the least common taxa of orthologous families (see Section 4.4) support the view that the root is located on the left in Figure 4.2. As expected the Fyn related kinase Frk and the proto-oncogene Abl sharing the SH2-SH3 domain combination cluster together. Further, the PNJ-tree strongly supports the view that the vascular endothelial growth factor receptors VEFGR, the fibroblast growth factor receptors FGFR, the proto-oncogene Ret and the protein tyrosine kinase Ryk/Dnt form a closely related group. The vascular endothelial growth factor receptor VEFGR-3 and the fibroblast growth factor receptor FGFR are traditionally grouped together. Both of them have Immunoglobulin-like receptor domains. While Hanks and Hunter [1995] state that the proto-oncogene Ret and the protein tyrosine kinase Ryk/Dnt have no close relatives, the tree of Robinson *et al.* [2000] weakly supports the mentioned relatedness. Yet, its internal edges are short and FGFR is closer to Ret than to VEFGR in the tree of Robinson. A drawback of the tree of Robinson *et al.* [2000] is the fact that it is constructed on human sequences only and that concerted evolution cannot be excluded. In contrast, profile distances of the PNJ tree are measured between species.

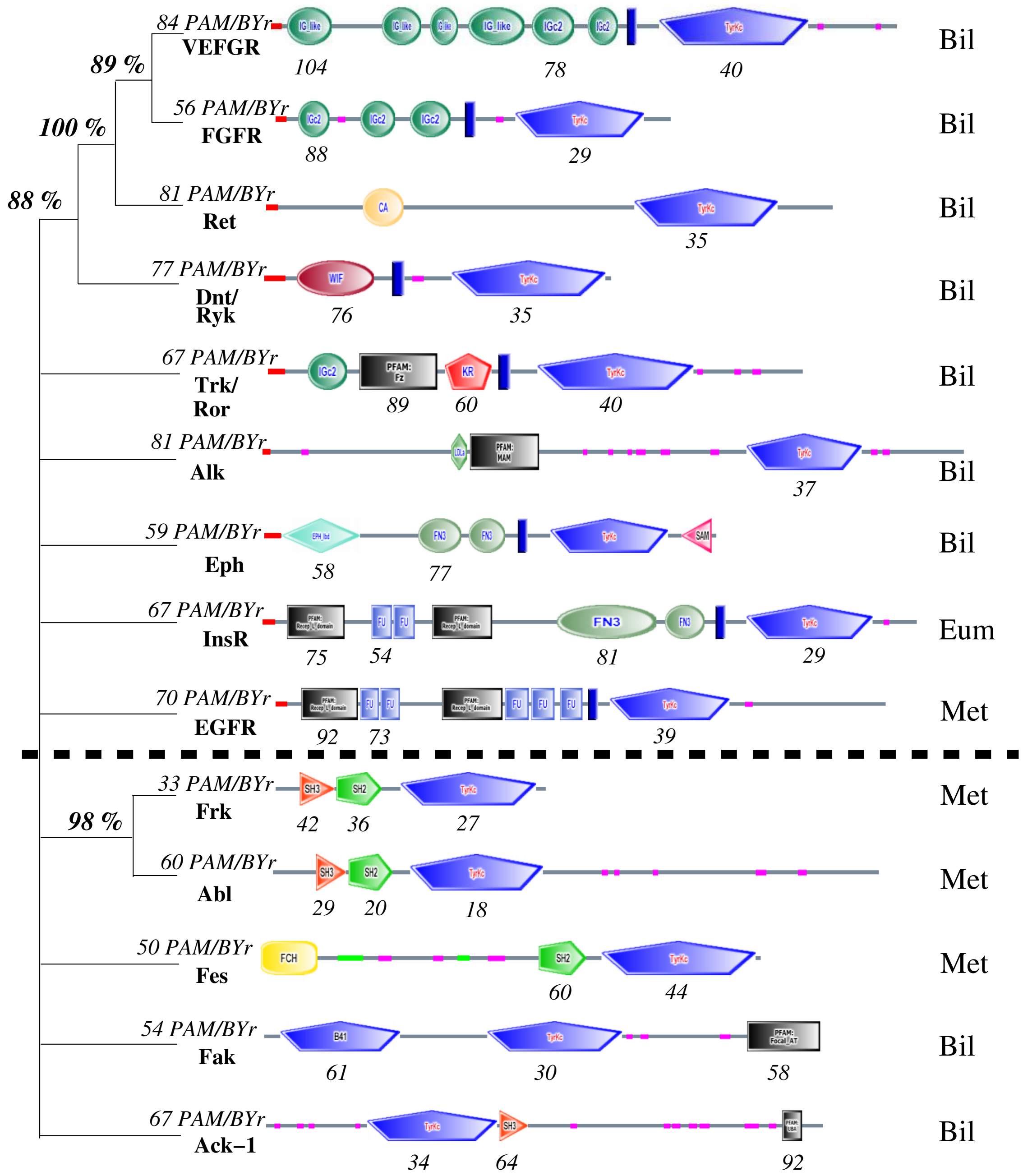


Figure 4.2: Receptor Tyrosine Kinases (above dashed line) and Protein Tyrosine Kinases (below dashed line). All domain architectures are present in SYSTERS cluster 136820 and within distinct orthologous families. Representations of the proteins by their domain architectures were downloaded from the SMART web server and comprise predicited SMART domains (colored bubbles) and PFAM domains (grey rectangles), transmembrane helices and signal peptides. Domain names are itemized in Appendix B. Leftmost, the PNJ-tree for the Tyrosine Kinase domains is depicted. Bold numbers at edges reflect the bootstrap support. FSRs of orthologous families are written above gene names. Numbers below domains are rates of domains in PAM/BYr units. The rightmost column lists the least common taxon of the orthologous families derived by domain architecture as described in Section 4.4.3 ("Bil" = Bilateria, "Eum" = Eumetazoa, "Met" = Metazoa).

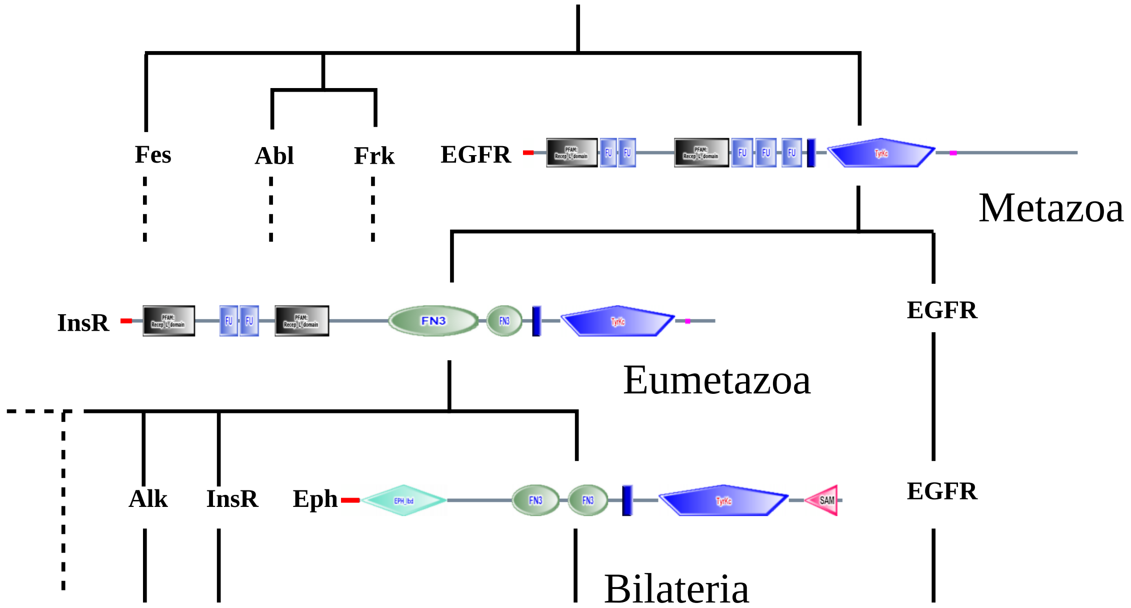


Figure 4.3: A parsimonious model for gene duplication and domain shuffling events during the evolution of receptor PTKs.

Domain shuffling

The mosaic like structured multidomain proteins within the PTK multigene family provide a prime example for the evolution of novel protein functions in Metazoa by duplication and domain shuffling events.

Consider the three rPTKs EGFR (Epidermal Growth Factor Receptor), InsR (Insulin receptor) and Eph (Ephrin tyrosine kinase receptor). The rightmost column in Figure 4.2 lists the least common taxon of the orthologous families reflecting the taxonomic range of the domain architecture (see Section 4.4). Arranging the domain architectures in a tree where internal nodes correspond to the least common taxa like in Figure 4.3 suggests a parsimonious model for a sequence of duplication and domain shuffling events: The domain architecture of EGFR comprising the Receptor L (Recep_L) domain, the Furin like repeats (FU) and the Tyrosine Kinase domain (TyrCk) already exists in the least common ancestor (LCA) of Metazoa. One round of gene duplication and domain shuffling occurs and a new protein evolves in the LCA of Eumetazoa by additionally acquiring the Fibronectin (FN) domain. Finally, Eph evolves in the least common ancestor of Bilateria by loosing the FU repeats and Recep_L and by acquiring the N-terminal Ephrin receptor ligand binding domain (EPH_lbd). (The sterile alpha motif domain (SAM) at the C-terminus of Eph is present in man and fugu. It is not detected in fly and an HMM-search of Eph in worm against the SAM model yields a weak hit with an E-value above the cutoff.)

Evolutionary rates

Family Specific Rates of orthologous families are written above gene names in Figure 4.2. While the rates of purely cytoplasmic PTKs range from 33 to 67 PAM/BYr, rates of receptor PTKs range from 56 to 84 PAM/BYr. This suggests that there is a general trend of the receptor PTKs to evolve at larger rates than the non-receptor PTKs.

We further disentangle the evolutionary rates by assessing the rates of the proteins' constituting domains. For that purpose we cut out domains detected in the sequences and align them to the domain models using "hmmalign". Finally, we apply the FSR estimator to the domain alignment. Domain rates are written below domain symbols in Figure 4.2. It is revealed that the extra-cellular domains are more divergent than their cytoplasmic counterparts. The largest rate observed for the Tyrosine Kinase domain is 44 PAM/BYr. In contrast, each of the extra-cellular domains is more divergent. We conclude that the large rates of the receptor PTKs indeed are due to the extra-cellular portions of the proteins.

4.3 Inferring the ancient origin of eukaryotic proteins

Having established that modern extra-cellular proteins are fast evolving, we examine the rates of orthologous families that have ancient origin. We trace the ancient origin of proteins by searching for similarities of orthologous families to proteins in prokaryotic organisms that imply a common ancestry. Figure 4.4 schematically illustrates the procedure. Alignments of the eukaryotic orthologous families serve as queries to perform a PSI-BLAST search against databases of complete prokaryotic organisms.

Protein sequences of 56 complete prokaryotic genomes were downloaded from the EBI [Apweiler *et al.*, 2004] and databases for each prokaryotic proteome were prepared. The complete list of prokaryotic organisms is given in Appendix C. Each orthologous family is searched against each prokaryotic genome by running one iteration of PSI-BLAST using default parameters. We consider only hits with an E-value below $E = 0.01$.

For each orthologous family, we count the prokaryotic organisms that are hit at least once with an E-value below the cutoff. The histogram is drawn in Figure 4.5. The black bar on the right corresponds to the set of 136 orthologous families that have hits in each of the 56 prokaryotic genomes. We call this set the "conserved" set of orthologous families. The grey bars hold the frequencies of 1801 orthologous families that do not have hits in all prokaryotic organisms but at least in four. We refer to this set as the "non-conserved" set. There is likely a considerable fraction of false positive hits of orthologous families that are found in less than four prokaryotic organisms. We label these families as "non-prokaryotic". The white bars in Figure 4.5 correspond to 1703

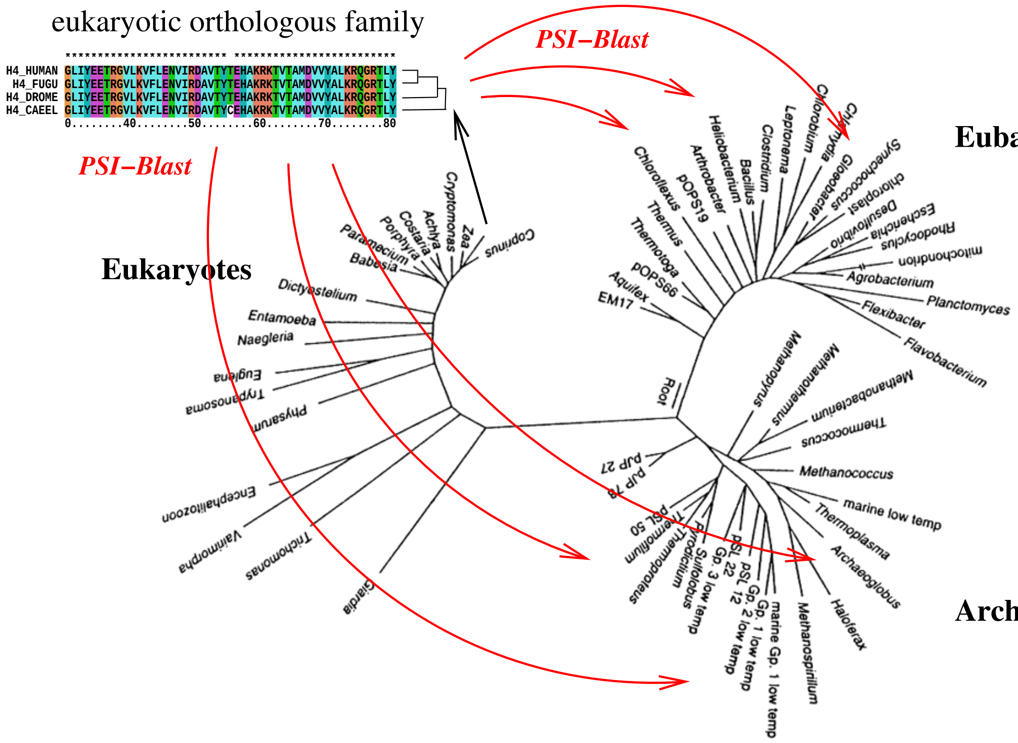


Figure 4.4: Schema of performed PSI-BLAST searches. Alignments of the eukaryotic orthologous families are interpreted as profiles and searched against databases of complete prokaryotic proteomes. The phylogenetic tree representation with species from all kingdoms of life is adapted from [Pace, 1997].

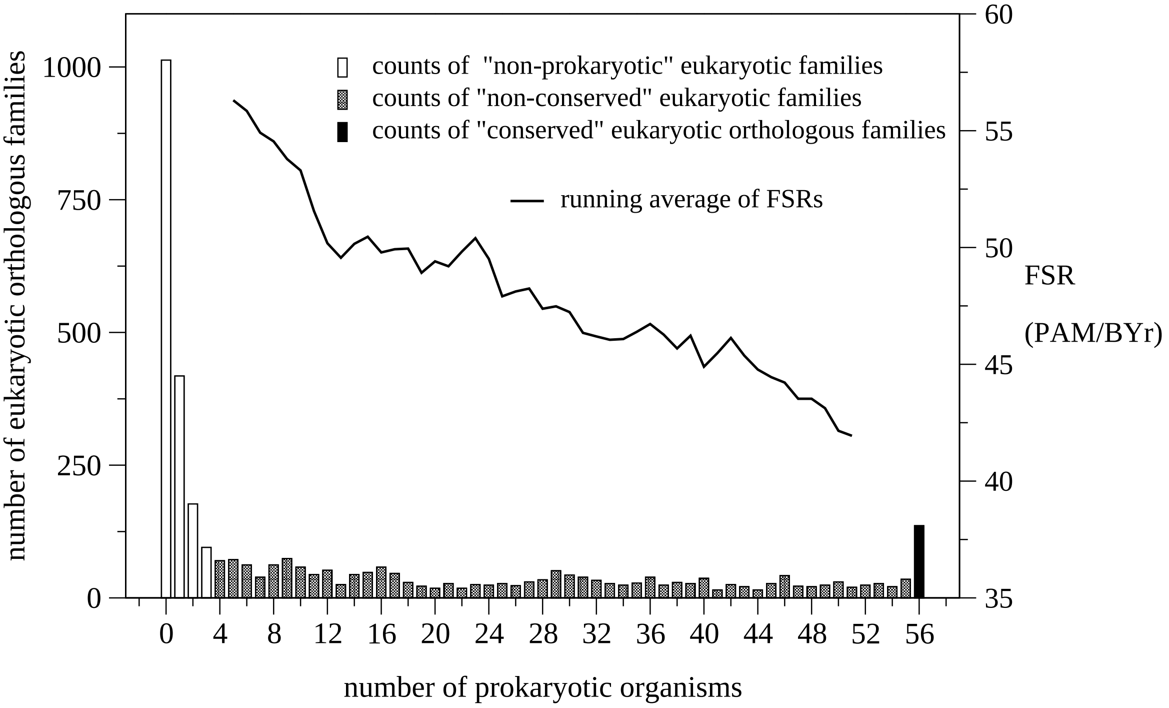


Figure 4.5: Counts of orthologous families that have hits in a certain number of prokaryotic organisms. The running average of Family Specific Rates over 11 bins respecitively is drawn as a line. The scale on the right refers to this line.

families of the "non-prokaryotic" set. All pairwise rate distributions are significantly different. Table 4.1 summarizes statistical values of the rate distributions.

subset	subset size	mean rate	standard deviation	median rate
all orthologous families	3640	52.4	25.1	50
conserved	136	35.0	18.9	33
non-conserved	1801	48.5	23.2	46
non-prokaryotic	1703	58.1	25.8	56
extra-cellular families	241	67.2	24.3	64

Table 4.1: Mean, standard deviation and median FSRs in the conserved, the non-conserved, the non-prokaryotic and the extra-cellular set in PAM/BYr units.

The conserved set is the set of orthologous families with the smallest mean FSR presented in this thesis. It includes ribosomal proteins that are known to be well conserved. Clearly, finding similarities of orthologous families within all prokaryotic organisms requires a high degree of conservation. Recall that the rates are measured in the metazoan branch of the tree (see Figure 4.4). An interesting point is that the proteins in the conserved set are supposed to have evolved slowly in common ancestors of eukaryotes and prokaryotes and still evolve at small rates within Bilateria.

Consider the conserved and the non-conserved set. First, orthologous families within both sets obey sequence similarities to prokaryotic sequences and thus are supposed to share a common evolutionary history with proteins of ancient origin. Second, the proteins of the non-conserved set are found in fewer prokaryotic organisms (or at larger E-value cutoffs, respectively) than the proteins of the conserved set. Third, the proteins of the non-conserved set evolve at larger rates than the proteins of the conserved set.

Three observations and a straight forward interpretation: The larger the rate of evolution, the less similarities are revealed by PSI-BLAST searches. The zigzag line in Figure 4.5 connects averaged FSRs for 11 bins respecitively. Its decreasing trend in the number of prokaryotic organisms affirms the above interpretation.

4.4 Taxon-specific rate distributions

4.4.1 Proteins evolved from primordial domains

PSI-BLAST searches allow to trace the ancient origin of almost 2000 eukaryotic orthologous families. Since PSI-BLAST reveals local similarities, this result does not

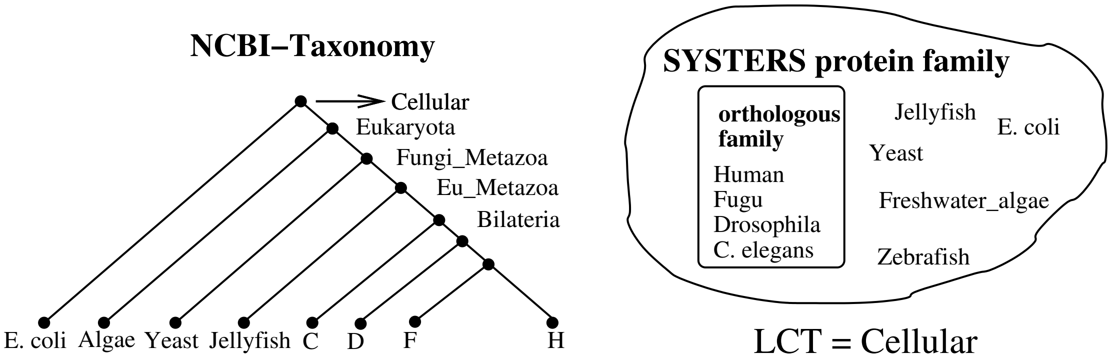


Figure 4.6: The least common taxon (LCT) with respect to the taxonomic range of a SYSTERS cluster. The orthologous family is mapped to a SYSTERS cluster. Members of the SYSTERS cluster are found in the eubacterium *E. coli*. According to the NCBI-taxonomy the LCT is *Cellular*.

imply that the proteins already existed in their current form in the common ancestor of eukaryotes and prokaryotes. Gene duplication followed by a recombination of protein domains is a fundamental and fast process that continually gives rise to novel proteins. While public sequencing lets the number of new proteins grow exponentially, the number of domains found seems to be close to saturation [Geer *et al.*, 2002]. There were likely no more than about 1000 protein domains in the primordial ancestors of present day organisms. The protein repertoire of present day organisms is supposed to have evolved from these domains [Chothia, 1994; Doolittle, 1995; Chothia *et al.*, 2003]. We proceed to assign an age to an orthologous family by its membership in a protein family. First we take the taxonomic range of all members of the protein family into account, second we consider proteins of the same domain architecture.

4.4.2 The least common taxon of a SYSTERS cluster

Within the SYSTERS database we are given a partitioning of the publicly available proteins into disjoint clusters representing protein families. 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.

The NCBI provides a curated taxonomy that classifies species by a hierarchical tree like structure with the aim of defining how species relate evolutionarily. To each orthologous family we assign a least common taxon (LCT) with respect to the NCBI-taxonomy and the members of the protein family it is mapped to. The procedure is sketched in Figure 4.6.

For example, an orthologous family which is labeled with the LCT *Cellular* is supposed to be of ancient origin while proteins of an orthologous family labeled with the LCT *Eukaryota* are supposed to be "younger". We call a set of orthologous families labeled with a specific LCT a *taxon-specific* set. As the sizes of the *Metazoa*- and the *Eumetazoa*-specific sets only amount to 33 and to 36 we merged these two sets into the composite *Eu_Metazoa*-specific set.

4.4.3 The least common taxon according to domain architecture

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 shown in Figure 4.2 including the Tyrosine Kinase domain all are present within one SYSTERS cluster. We think that a protein that takes a well defined function and has a specific age obeyes a specific domain architecture.

In order to assess domain architectures we search the sequences of the orthologous families and the sequences of SYSTERS clusters for SMART and Pfam domains using "hmmpfam", 7316 Pfam-HMMs from Release 12.0 and 662 SMART-HMMs from Release 3.7. A domain architecture is defined as a sequence of domain occurences where we count multiple consecutive occurences of the same domain as one occurrence only. 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 cover 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 show up a unique domain architecture. In this way 3177 orthologous families are labeled with a domain architecture.

For a given orthologous family and its domain architecture we group together homologous sequences within the respective SYSTERS cluster that have the same domain architecture. To the orthologous family a least common taxon is assigned with respect to the NCBI-taxonomy and to the sequences in such a group. Figure 4.7 gives an example for a multigene family and two orthologous families X and Y: When considering the taxonomic range of the sequences in the SYSTERS cluster only, both orthologous families are labeled with the same LCT *Cellular*. Instead, the SYSTERS cluster is subclustered and the LCT *Eu_Metazoa* is assigned to orthologous family Y.

4.4.4 Taxon-specific rate distributions

We compare the rate distributions of taxon-specific sets. Figure 4.8 shows the cumulative normalized histograms of Family Specific Rates when assigning the LCT of

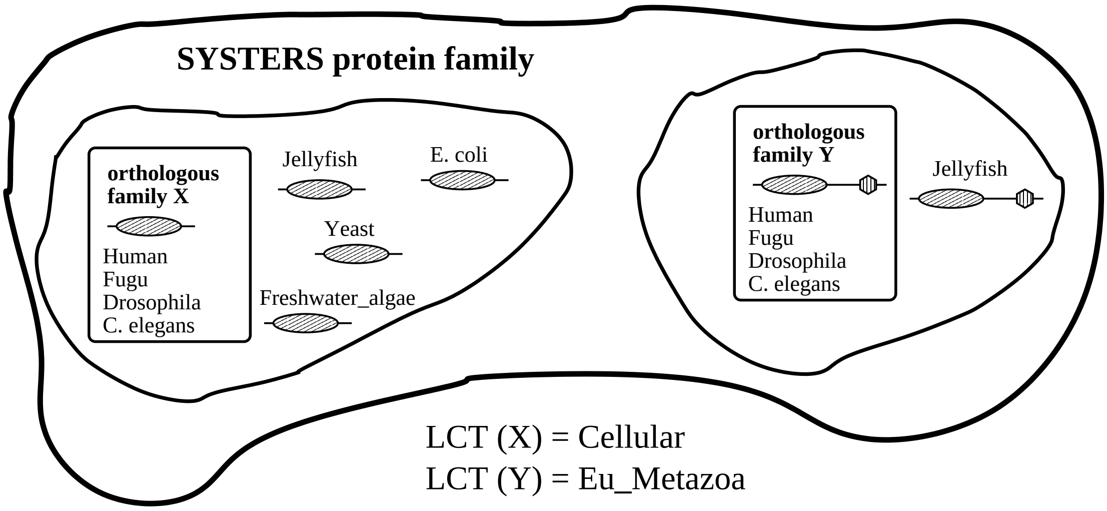


Figure 4.7: The least common taxon according to domain architecture. Two orthologous families are mapped to one SYSTERS cluster representing a multigene family. The LCT for each orthologous family is assessed by considering the sequences in the SYSTERS cluster having the same domain architectures.

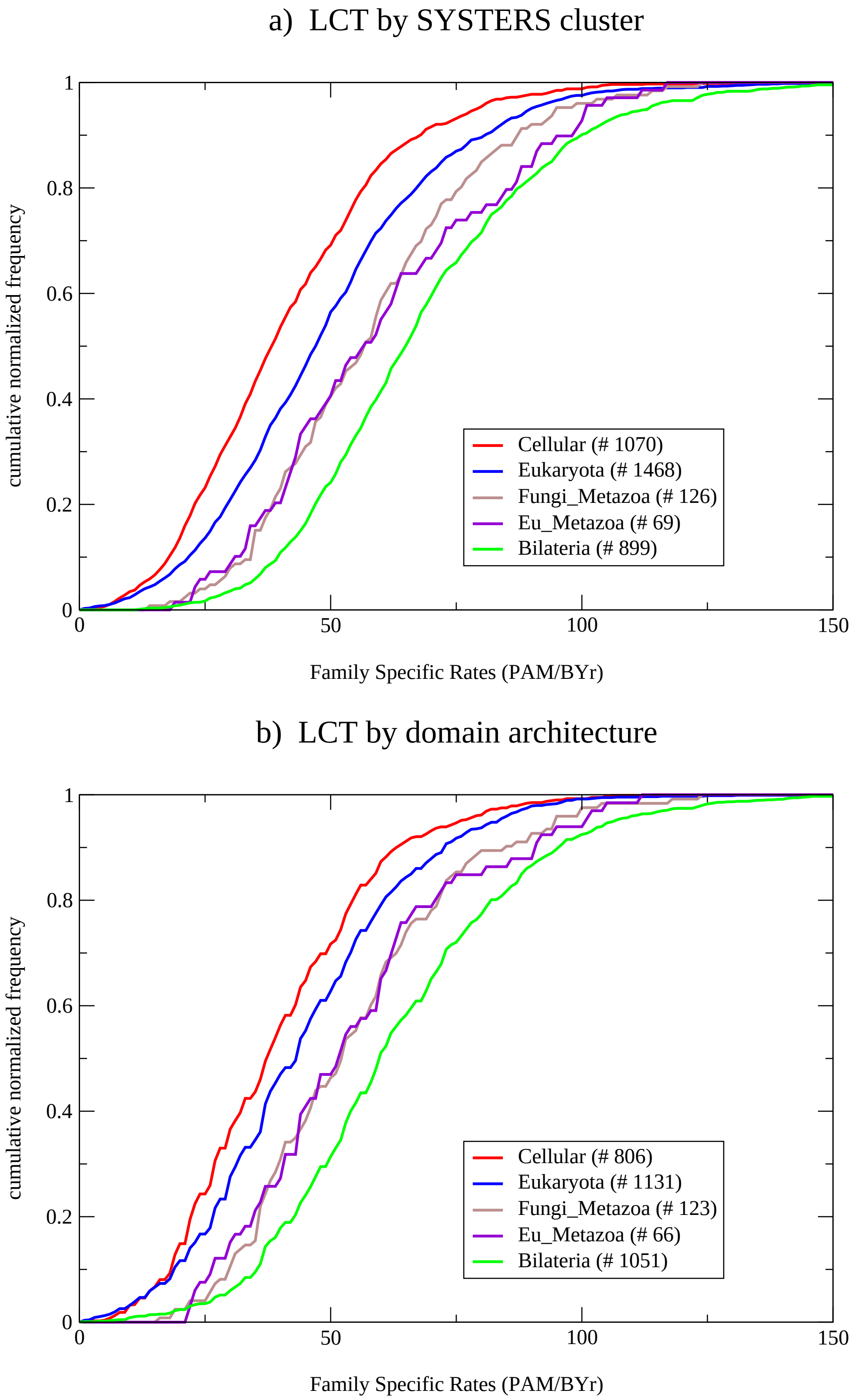
an orthologous family by considering all sequences in a SYSTERS cluster (Figure 4.8 a) and by considering the sequences in the SYSTERS cluster with the domain architecture of the orthologous family (Figure 4.8 b). The results are summarized in Table 4.2.

We observe that the mean rates and the cumulative rate distributions 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 taxon-specific rate distributions are significant. The most divergent orthologous families with no detectable domains are missing in taxon-specific sets that were derived by domain architecture. As a consequence, the mean Family Specific Rates of those sets are smaller.

In the following sections we refer to the taxon-specific sets that were derived by domain architecture.

4.4.5 Metazoan radiation and the invention of extra-cellular proteins

The taxon-specific sets facilitate to shed light on the emergence of proteins during the metazoan radiation. We consider intersections of taxon-specific sets with the set of extra-cellular orthologous families. In Figure 4.9 b) the height of a bar holds the relative number of extra-cellular orthologous families within a taxon-specific set and



a) LCT by SYSTERS cluster

b) LCT by domain architecture

Figure 4.8: Normalized cumulative histograms of taxon-specific rates. The least common ancestor of an orthologous family is assessed by: a) all sequences being present in a SYSTERS cluster b) the sequences in a SYSTERS cluster having the same domain architecture.

	LCT by SYSTERS cluster				LCT by domain architecture			
taxon-specific subset	subset size	mean FSR	std dev	*p*-value	subset size	mean FSR	std dev	*p*-value
Cellular	1070	41.5	20.4	–	806	39.7	19.1	–
Eukaryota	1468	49.5	23.2	1e–19	1131	44.4	21.0	3e–7
Fungi_Metazoa	126	58.4	21.8	1e–5	123	54.3	21.5	3e–6
Eu_Metazoa	69	60.5	24.4	0.679	66	54.3	22.5	0.919
Bilateria	899	68.2	24.5	0.011	1051	63.0	24.7	0.003

Table 4.2: Taxon-specific rate distributions. Mean and standard deviation are given in PAM/BYr. *p*-values stem from pairwise Wilcoxon two sample test where the set in the row of the *p*-value was compared to the set one row above.

the surface area of a bar corresponds to the absolute size of an intersection, that is to the total number of extra-cellular families within a taxon-specific set. The bars peak for *Eu_Metazoa*, 19 of 66 *Eu_Metazoa*-specific families and 155 of 1051 *Bilateria*-specific families belong to the set of extra-cellular families. Thus, the proposition that extra-cellular proteins mainly were invented during metazoan evolution is approved.

4.4.6 The younger the faster?

Taxon-specific rate distributions are in accordance with the hypothesis "the younger the faster". Clearly, the rate distributions are subject to a systematic influence. We have already seen that our ability to detect homology by sequence comparison is restricted by the amount of similarity the sequences obey (see Section 4.3). 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.

The set of fast evolving extra-cellular proteins was obtained without considering taxon-specificity of the proteins at all. Thus, since the taxon-specific sets accentuate the modernity of extra-cellular proteins, "the younger the faster" is supported. The latter is a circular statement. Because, under the assumption that extra-cellular proteins *per se* evolve fast, the taxon-specific rate distributions are expected to support the view of extra-cellular proteins having emerged relatively recently. Vice versa, in Section 3.6.6 we have seen that enzymes evolve at small rates. Relative sizes of intersections of enzymes to taxon-specific sets are shown in Figure 4.9 c. Indeed, the enzymes account for almost 50% of the proteins in the *Cellular*-specific set.

Still, indirectly assessing homology relations by domain architectures is benefitial with respect to homology detection since domains constitute the conserved parts of a protein

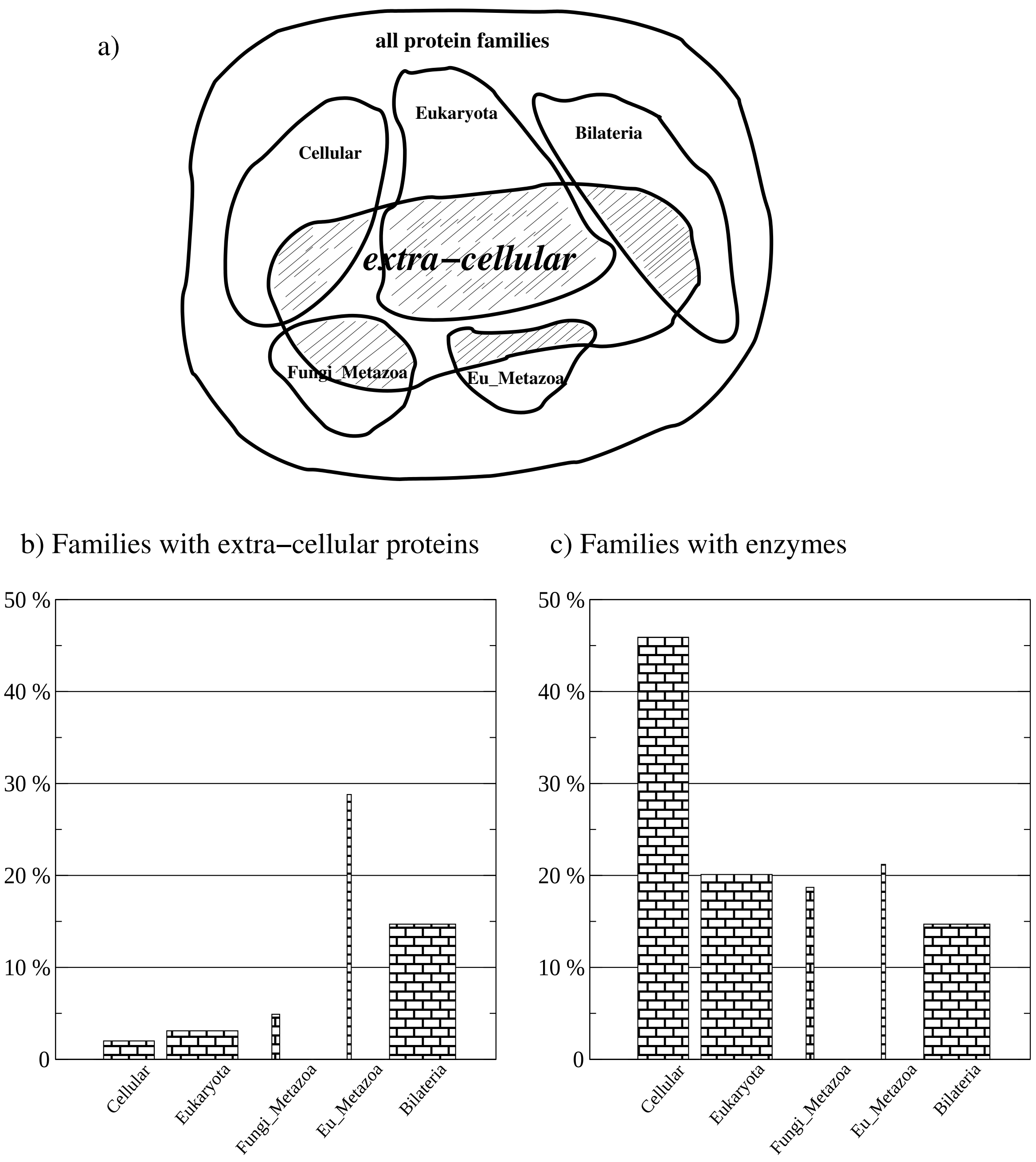


Figure 4.9: The emergence of extra-cellular proteins and enzymes. a) The taxon-specific sets and the set of extra-cellular families overlap. The height of a bar in b) is the size of an intersection divided by the size of the taxon-specific set. Bar widths correspond to sizes of the taxon-specific sets. c) Intersections of taxon-specific sets with the set of enzymes.

and searching domains with Hidden Markov Models has been proven to be powerful and sensitive. The taxonomic range of an orthologous family that is assessed by all sequences in a SYSTERS cluster is larger or equal to the taxonomic range assessed by domain architecture. Interestingly the differences between taxon specific rate distributions are in the same range.

Further consider the *Fungi_Metazoa*-, the *Eu_Metazoa*- and the *Bilateria*-specific sets. Altogether these sets comprise 1240 orthologous families, but the fraction of slowly evolving proteins in these sets is small (see Figure 4.8). The dependance of our means to detect homologies on evolutionary rates is irrelevant regarding this fact.

4.5 Multigene families

4.5.1 Duplicated genes are more conserved

We investigate multigene families and estimate time points of duplication events predating the nematode-arthropode split that gave rise to taxon-specific orthologous families. Since we want to obtain multiple alignments of orthologous families to measure evolutionary distances among paralogs, we define a multigene family as a group of orthologous families sharing the same domain architecture.

First, we split the set of all orthologous families into a *singleton* and into a *duplicate* set. While the singleton set comprises 1689 orthologous families with a domain architecture being unique for the respective family, the duplicate set contains 1488 orthologous families with domain architectures that are present in at least two orthologous families.

Figure 4.10 shows normalized cumulative rate distributions of the duplicate and the singleton set. The rate distribution of the singelton set is significantly shifted to larger rates. The p-value of a Wilcoxon two sample test is $p = 1.72 \cdot 10^{-9}$. This confirms the results of two studies which aimed at showing that genes being in general prone to duplications are more evolutionarily conserved than genes with no detectable copy [Davis and Petrov, 2004; Jordan *et al.*, 2004]. Still, if we fail to detect paralogies for fast evolving families, the set of singletons is biased towards larger rates.

4.5.2 “Young” and “old” multigene families

“Young” and “old” multigene families

We further investigate the duplicate set and define a multigene family as a set of orthologous families that obey the same domain architecture. We obtain 433 multigene families that are made up of 1696 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 taxon specificity of the orthologous families. While the size of the young set amounts to 95 the old set comprises 338 multigene families.

Paralogous evolutionary profile distances

We interprete the alignments of orthologous families as orthologous profiles. For each pair of orthologous families within a multigene family we compute a profile alignment using CLUSTALW [Thompson *et al.*, 1994; Higgins *et al.*, 1996] with default parameters. A profile alignment in turn serves to compute the evolutionary profile distance [Müller *et al.*, 2004] using the Müller-Vingron model. The evolutionary profile distance t_{XY} holds the average number of substitutions which have accumulated according to the model between any two paralogous sequences of orthologous families X and Y.

Figure 4.11 shows scatter plots comparing all pairwise profile distances and Family Specific Rates in the “young” set $\mathcal{Y}$ and in the “old” set $\mathcal{O}$ of multigene families. Family Specific Rates are averaged for the two orthologous families compared. Within $\mathcal{O}$ the data points of the 8 largest multigene families with more than 20 members are discarded. Big circles indicate the center of mass or the mean of the data point positions. For $\mathcal{Y}$, the center of mass is located at (63 PAM/BYr, 258 PAM), for $\mathcal{O}$ it is at (44 PAM/BYr, 248 PAM). The two ratios of mean paralogous distances and rates reflect the range of averaged duplication times. The smaller ratio in the scatter plot for “young” multigene families points to the possibility that duplications occured more recently within $\mathcal{Y}$ than within $\mathcal{O}$.

Further, we compare levels of sequence similarity among orthologous families (that are paralogous) to Family Specific Rates by letting each multigene family contribute only once to the comparison: Evolutionary rates between $\mathcal{Y}$ and $\mathcal{O}$ are compared by assigning to each multigene family i the arithmetic mean $\bar{\lambda}_i$ of its Family Specific Rates. Similarily, we average all pairwise profile distances within multigene family i and obtain $\bar{t}_i$. Table 4.3 gives the results.

In accord with the cumulative rate distributions of taxon-specific sets, 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 $2.67 \cdot 10^{-10}$. And 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.

Interestingly the same does not hold when comparing mean paralogous profile distances. 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/My.

	#	average evolutionary FS 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
ranksum p-value		$2.67 \cdot 10^{-10}$	0.717

Table 4.3: Rates and paralog distances in young and old multigene families The table compares rates and paralogous evolutionary distances in the two sets $\mathcal{Y}$ and $\mathcal{O}$ of young and old multigene families.

Genes in multigene families with a "younger" least common taxon were duplicated more recently

Figure 4.12 schematically compares two multigene families, one of $\mathcal{Y}$ and one of $\mathcal{O}$: While 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" family. Here the vertical axis scales with evolutionary distances. Then, interpreting the situation with respect to a time scale, the duplication time points within the "old" family are larger than the duplication time points for the "young" family. Thus, an argument for the pertinence of assigning an age to a protein by the taxonomic range in the protein family is placed.

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 family of $\mathcal{Y}$ amounts to 3.1 and to 4.1 in a multigene family of $\mathcal{O}$. We repeat the analysis and exclude the 8 largest multigene families with more than 20 orthologous families such that the average number of orthologous families in $\mathcal{O}$ decreases to 3.3. The results presented in Table 4.3 practically do not change. For instance, the average rate remains the same, 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 $2.67 \cdot 10^{-10}$.

Further, we estimate the time points of early duplication events for pairs of orthologous families. The procedure requires a relative rate test that involves a third distantly related orthologous family as "outgroup". There are no outgroup families for the most distantly related pairs of orthologous families. That is, when estimating duplication time points, we implicitely set the most distant relationships aside.

4.5.3 Dating duplication events

Dating duplication events

The putative time point of a duplication event pinpoints the time of protein emergence and indicates the age of the proteins. Consider two orthologous families X and Y of such a 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 predating the arthropode-nematode split. Given paralogous relationships between sequences only, one has no clue of how to assess the timepoint of a duplication event being unique to the given family. Yet, under the assumption of rate constancy it is possible to extrapolate the rates measured on homologous regions between orthologs to infer the time of the duplication event. Rate constancy is tested by a *relative rate test* [Sarich and Wilson, 1973]. The relative rate test requires that there is a third more distantly related outgroup family Z present in the multigene family (see Figure 4.13). We require that the LCT of Z is the same or a predecessor of the LCTs of X and of Y.

We use CLUSTALW to first align X and Y and then Z to the profile alignment of X and Y. We call gapless subalignments of the homologous regions in X, Y and Z that comprise orthologous sequences *orthologous profiles* x, y and z (see Figure 4.13). The Family Specific Rate measure is applied to assess the profile rates λ_x and λ_y corresponding to orthologous profiles x and y. Under the assumption that λ_x and λ_y are constant in X and Y, we set $\lambda_{xy} = (\lambda_x + \lambda_y)/2$. The estimate of the duplication time τ_{xy} is obtained from the relation

$$t_{xy} = 2 \cdot \lambda_{xy} \cdot \tau_{xy}$$

where t_{xy} is the evolutionary profile distance between orthologous profiles x and y.

The rate extrapolation is based on the assumption that the rate of protein evolution on homologous regions of the duplicated proteins remained approximately constant following the duplication as well as 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 λ_x and λ_y differed by a factor larger than 4/3. Rate constancy subsequent to the duplication event is tested by the relative rate test.

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 4.13). 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, $\widehat{\delta t} = 0$ is practically not observed. To assess the uncertanties of the profile distance measures, 100 bootstrap replicates of the profile alignment of X, Y and 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 $\widehat{\delta t}$ from 0 is in the size of the measurement error, the probabilities that $\widehat{\delta t} < 0$ or that $\widehat{\delta t} > 0$ 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 $\widehat{\delta t} < 0$ and $\widehat{\delta t} > 0$ are sampled from a binomial distribution with $p_0 = \Pr(\widehat{\delta t} < 0) = \Pr(\widehat{\delta t} > 0) = 0.5$. The relative rate test is passed when H_0 can be accepted at a significance level of 95%, that is, when the number of times where $\widehat{\delta t} < 0$ for 100 bootstrap replicates is larger than 40 and smaller than 60.

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 4.14 compares the duplication times to the profile rates. The profile rates λ_{xy} are close to the Family Specific Rates $\hat{\lambda}_X$ and $\hat{\lambda}_Y$. Since they were calibrated by using the same set of divergence times that were used to calibrate the Family Specific Rates, we do not observe instances where the estimated duplication times are significantly smaller than 1170 Millions of years, the given time for the arthropode-nematode split. Still, there are 18 estimated times exceeding 4 billions of years, the putative age of the earth, and the estimated duplication times in general are large. This might be caused by overestimated divergence times [Peterson *et al.*, 2004; Graur and Martin, 2004] or a systematic decrease in evolutionary rates following the duplication event.

The younger the faster

The scatter plot in Figure 4.14 is in accordance to "the younger the faster". While duplication times of slowly evolving proteins cover a wide range, the upper bound of estimated duplication times decreases with increasing rates.

We further draw the connection of estimated duplication times to the taxon-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. In accordance to 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 ranksum test yields a p-value of $p = 0.14$. Yet the estimated duplication times for the old set are significantly larger than for the young set ($p = 1.80 \cdot 10^{-6}$).

The estimated duplication times support the view that genes in multigene families with a small taxonomic range were duplicated more recently than genes in multigene families with a broad taxonomic range.

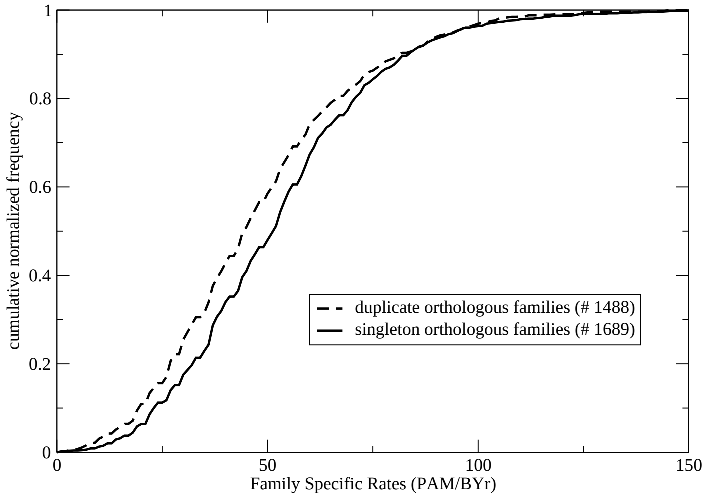


Figure 4.10: Cumulative normalized rate distributions of the singleton set and the duplicate set of orthologous families. The duplicate set contains orthologous families that occur in more than one orthologous family. The domain architectures in the singleton set are only found once.

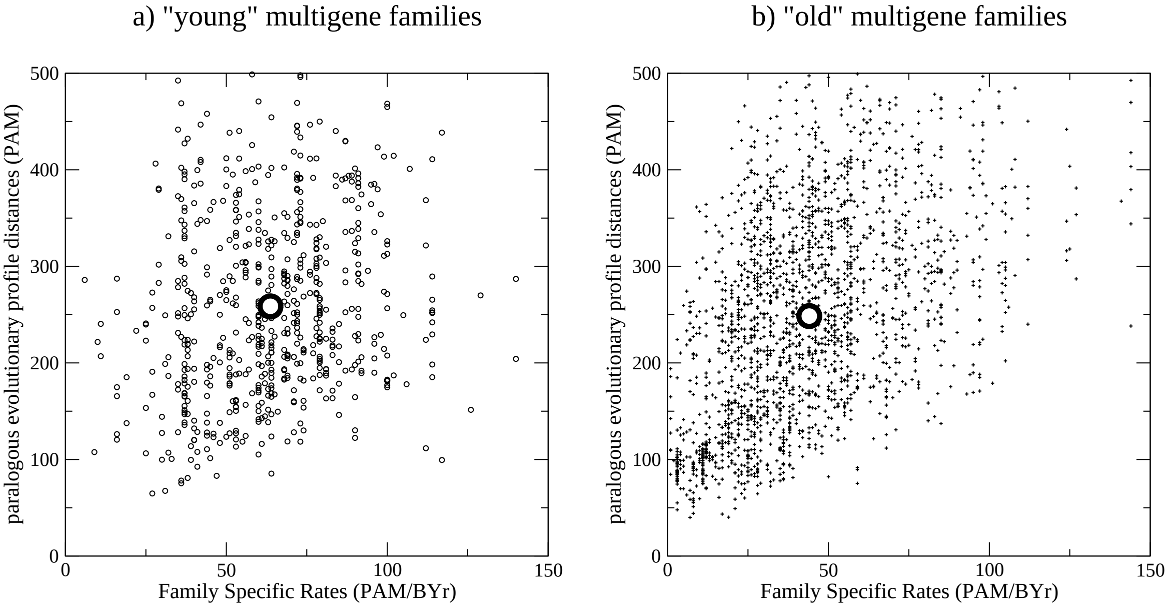


Figure 4.11: Scatter plots comparing paralogous evolutionary distances and Family Specific Rates for multigene families with a ”young” LCT (*Bilateria*, *Eu_Metazoa* or *Fungi_Metazoa*) (a) and with an “old” LCT (*Eukaryota* or *Cellular*) (b). Family Specific Rates were averaged for two orthologous families respectively. Big circles indicate the mean of the data points’ positions.

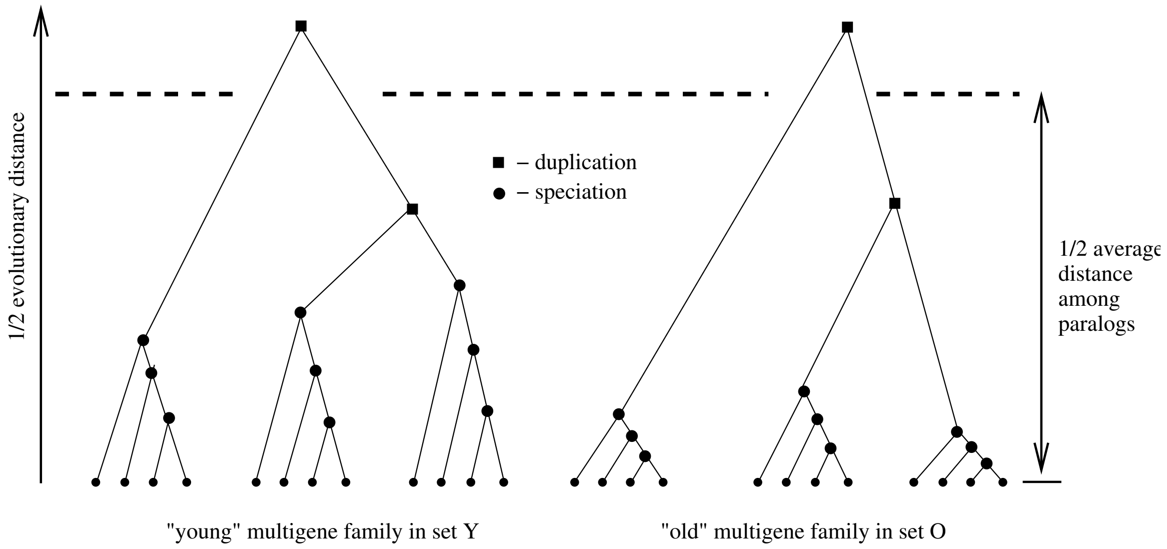
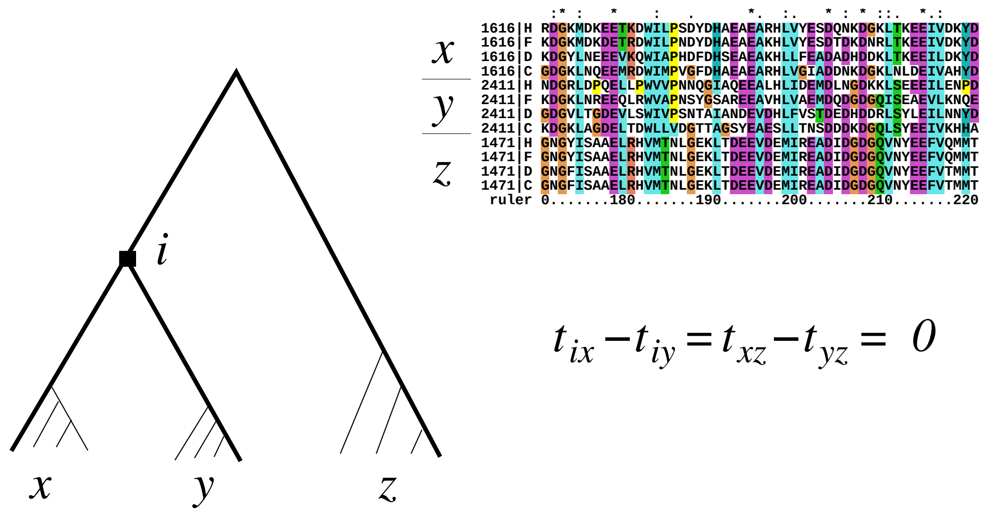

1/2 evolutionary distance
■ – duplication
● – speciation
1/2 average distance among paralogs
"young" multigene family in set Y
"old" multigene family in set O

Figure 4.12: Schematic representation of multigene families. On the left the tree for a multigene family of the "young" set $\mathcal{Y}$ is shown. The tree on right represents a multigene family of the "old" set $\mathcal{O}$. 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.



```
        :* :   *    :   .      *.  :.   * : * ::.  *.:
1616|H RDGKMDKEETKDWILPSDYDHAEAEARHLVYESDQNKDGKLTKEEIVDKYD
1616|F KDGKMDKDETRDWILPNDYDHAEAEAKHLVYESDTDKDNRLTKEEIVDKYD
1616|D KDGYLNEEEVKQWIAPHDFDHSEAEAKHLLFEADADHDDKLTKEEILDKYD
1616|C GDGKLNQEEMRDWIMPVGFDHAEAEARHLVGIADDNKDGKLNLDEIVAHYD
2411|H NDGRLDPQELLPWVVPNNQGIAQEEALHLIDEMDLNGDKKLSEEEILENPD
2411|F KDGKLNREEQLRWVAPNSYGSAREEAVHLVAEMDQDGDGQISEAEVLKNQE
2411|D GDGVLTGDEVLSWIVPSNTAIANDEVDHLFVSTDEDHDDRLSYLEILNNYD
2411|C KDGKLAGDELTDWLLVDGTTAGSYEAESLLTNSDDDKDGQLSYEEIVKHHA
1471|H GNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT
1471|F GNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT
1471|D GNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMT
1471|C GNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMT
 ruler 0.......180.......190.......200.......210.......220
```

Figure 4.13: 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.

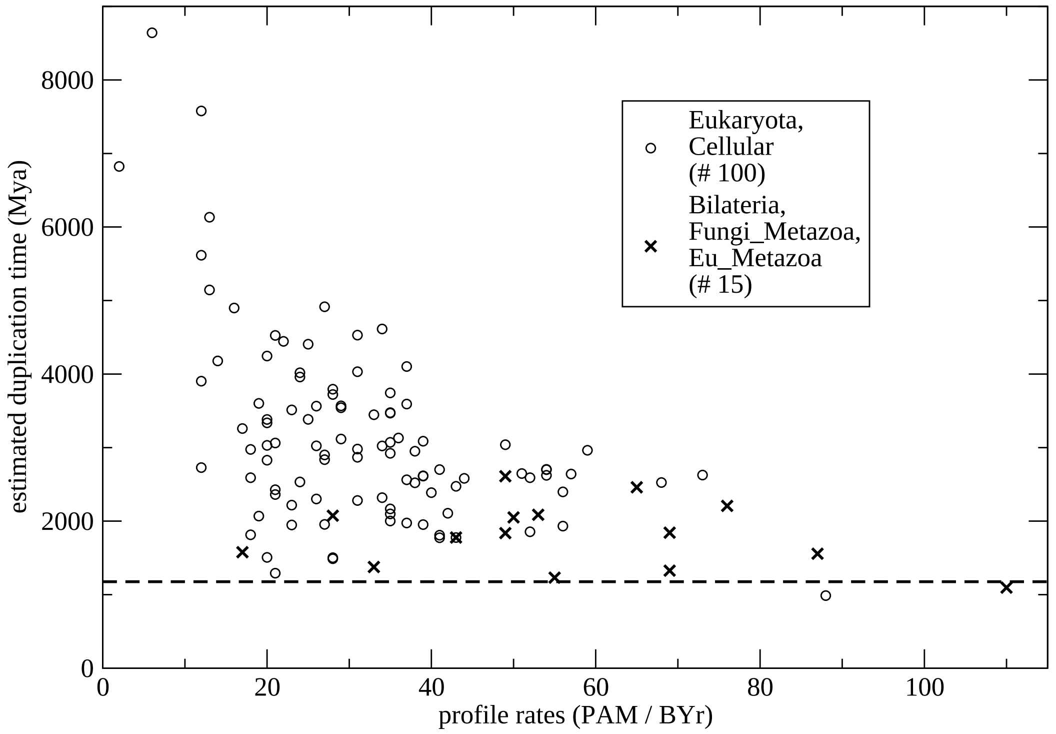


Figure 4.14: Estimated duplication time points compared to profile rates measured on homologous regions of profile alignments between orthologous families. Circles hold times estimated for orthologous families with "old" LCT, crosses mark times estimated for families with a "young" least common taxon. The dashed line is placed at 1170 Millions of years as the time of the nematode-arthropode split that we used to calibrate Family Specific Rates.