3 Family specific rates of protein evolution

3.1 Rates of protein evolution

The rates of protein evolution are routinely quantified by comparing the coding nucleotide sequences of orthologous gene pairs between closely related organisms. Several authors apply the codon substitution model implemented in PAML [Yang, 1997] to estimate d_N, the expected number of non-synonymous substitutions causing a change of the amino acid sequence [Stein *et al.*, 2003; Davis and Petrov, 2004; Jordan *et al.*, 2004; Castillo-Davis *et al.*, 2004]. The molecular clock model assumes that the number of substitutions is proportional to the divergence time and to a constant rate at which substitutions accumulate. Since orthologous sequences have diverged by speciation, the obtained values of d_N constitute the rate distribution of the proteomes. Still when d_N, measured between two close lineages, is transferred to other distantly related protein coding genes as in [Davis and Petrov, 2004], rate variations among diverse lineages are not taken into account. Here it becomes feasible to estimate evolutionary rates by measuring the degree of sequence divergence among sets of orthologous proteins or *orthologous families*. Requiring the orthologous families to hold members of the same organisms ensures that the time that has passed since the sequences diverged is constant within different orthologous families. Thus different levels of sequence divergence can be compared and related to historical time. For example Koonin *et al.* apply a measure for an evolutionary rate on sets of distantly related orthologs by averaging distances from the outgroup sequence to other sequences [Koonin *et al.*, 2004]. We apply standard concepts to compute maximum likelihood trees which in turn provide measures for a family specific evolutionary rate. Considering evolutionary paths in a phylogenetic tree over a large time scale has the major effect of averaging out lineage specific rate variations.

The promising goals when pinpointing rate distributions include the disclosure of global principles influencing selection. A general procedure is to search for a biologically significant partitioning of protein sets obeying different rate distributions. Suggested criteria to do so tend to relate some degree of protein dispensability, measured for instance by RNA interference or the number of interaction partners, to their rates [Jordan *et al.*, 2002; Hirsh and Fraser, 2001; Fraser *et al.*, 2003; Teichmann,

2002]. The rationale is that purifying selection is expected to act weaker on essential genes than on nonessential ones. Although these studies are appealing, correlations are sometimes weak and interpreting the results is subject to controversial discussions. Other authors accomplish correlating evolutionary rates with sequence length, tissue specificity or the affiliation of the proteins to functional categories [Lipman *et al.*, 2002; Winter *et al.*, 2004; Castillo-Davis *et al.*, 2004] .

The sections of this chapter are organized as follows. Sections 3.2 and 3.3 outline the preparation of orthologous families with representatives from man, fugu, fly and worm and motivate the choice of the substitution model. Section 3.4 compares evolutionary distances between pairs of organisms. The presented scatter plots provide empirical evidence for the assertion that a family specific selective pressure constitutes the dominant effect influencing the rates of protein evolution. We suggest and apply two ML estimators for a family specific evolutionary rate (Section 3.5). In Section 3.6 rate distributions are presented. We compare the family specific rates to measures of nonsynonymous substitutions between *C. elegans* and *C. briggsae*. Further, we make use of published data of RNAi- as well as of two-hybrid experiments in the nematode model and establish that essential genes tend to be evolutionary more conserved than others. Finally, rate distributions of orthologous families occupying certain functional classes are examined.

3.2 The data, orthologous families and alignments

We derive orthologous families containing members of the primate *Homo sapiens*, the pufferfish *Fugu rubripes*, the arthropode *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*. The sample among completely sequenced and divergent model organisms is chosen such that pairs of orthologous amino acid sequences are subject to a significant and informative portion of sequence divergence. On the other hand the selection of organisms allows to produce a large number of global multiple alignments since the orthologs mainly diverged on the sequence level and not, e.g., by domain shuffling. The peptide sequences were downloaded from the *Ensembl* database (version 16, September 2003) [Hubbard *et al.*, 2002].

We first apply the INPARANOID software (see Section 2.3.5) to obtain orthologous groups for each pair of organisms by requiring a large confidence for orthologous assignments and setting the INPARANOID confidence value to 95%. Under the assumption that orthologous relationships are transitive, the orthologous groups derived for pairs of organisms are merged into orthologous families if they have a sequence in common. Altogether $10,769$ orthologous families are derived. Among those, $3,992$ families containing at least one representative of each organism are further analysed and put into the multiple alignment procedure (see Figure 3.1).

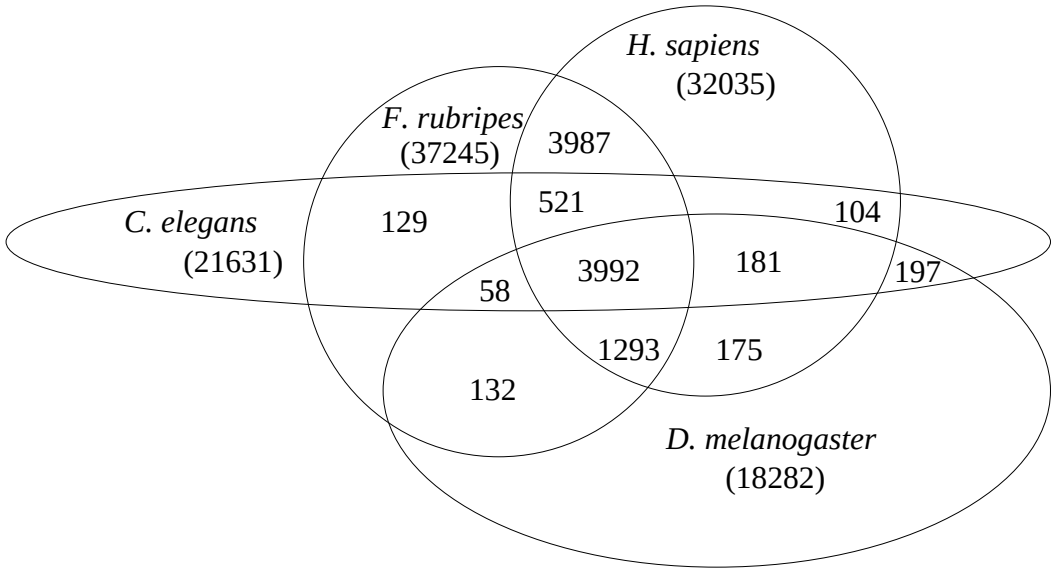

Figure 3.1: The distribution of species among orthologous families. Numbers in parentheses indicate the number of sequences put into the clustering procedure. Numbers within intersections indicate the number of orthologous families containing representatives of the intersecting species.

For each of the $3,992$ families four sequences from different species are selected as orthologous representatives of the family for the multiple alignment. Some orthologous families contain more than one representative per organism. First and by definition of in-paralogy, recent gene duplications are expected to have occured. The possibility that concerted evolution has acted cannot be excluded, but the large confidence value chosen for in-paralogous assignments is supposed to reduce this effect. Second, the database of peptides contains different splice variants of the same gene that share identical sequence regions.

Thus, it is desirable to multiply align sequences having similar lengths. When more than one sequence per organism is present in an orthologous family, we align four sequences by requiring that their sequence lengths are similar. To be concrete, we consider all sets of four sequences from different organisms and choose the one that has the minimal sum over quadratic length differences

$$\Delta L = \sum_{i \neq j} \frac{(l_i - l_j)^2}{(\min(l_i, l_j))^2} \ ,$$

where l_i is the length of a sequence in organism i. In order to penalize occurences of sequences having small lengths as a consequence of missing or unrecognized exons, weighting factors are chosen as reciprocal squared lengths of the small sequences.

Prior to multiply aligning the sequences, they are filtered for low complexity regions using SEG [Wootton and Federhen, 1993]. For each family a multiple alignment of four orthologous sequences is generated using DCA. The recursion stop size in DCA is set to 400. That is, obtained multiple alignments with less than 400 sites are definite optimal alignments with respect to the sum of pairs score. In other respects we choose default parameters of DCA including BLOSUM62 as scoring matrix and free end gaps (gaps occuring at the beginning or the end of the alignment were not penalized). Finally we discard orthologous families with alignments containing less than 80 gapless sites as well as some families with spurious alignments containing large numbers of gaps. Altogether we end up with a set of $3,640$ orthologous families and multiple alignments.

3.3 The substitution model

The codon substitution model is commonly used to estimate the number d_N of nonsynonymous substitutions between orthologous coding nucleotide sequences of closely related organisms. Since the codon substitution model implies equal distances for any pair of amino acids, it is simplifying with respect to the process of amino acid replacement. Further, nonsynonymous substitutions over a large time scale are subject to saturation.

Recall that our focus is the degree of sequence divergence among orthologs. The input data for estimating the replacement frequencies in the Müller-Vingron model were alignments of varying degree of divergence (see Section 2.4.5). We therefore analyze the alignments directly on the amino acid level and choose the Müller-Vingron model derived by the ML approach as amino acid replacement model [Müller *et al.*, 2002]. ML evolutionary distances and phylogenetic trees are computed on the gapless sites of the alignments.

3.4 Comparing pairwise evolutionary distances

We estimate evolutionary distances in PAM units. Evolutionary distances between orthologs of two organisms reflect a rate distribution. For four organisms there are $6 = \binom{4}{2}$ pairs of organisms. Within orthologous families, we compare evolutionary distances between two different pairs of organisms and produce $15 = \binom{6}{2}$ scatter plots. For example, the distances that orthologous pairs of man (H) and fugu (F) have are compared to the distances of the corresponding orthologous pairs of fly (D) and worm (C).

	HD	HC	FD	FC	DC
HF	0.634	0.556	0.695	0.597	0.460
HD		0.784	0.922	0.750	0.795
HC			0.742	0.943	0.905
FD				0.791	0.781
FC					0.890

Table 3.1: The table shows correlation coefficients when comparing two vectors of 3640 evolutionary distances respectively (see Figure 3.2).

Figure 3.2 shows orthologous distances between H and C on the one hand and between F and C on the other hand (a), as well as distances between H and F and between D and C (b). We call these scatter plots HC-FC scatter plot and HF-DC scatter plot. All scatter plots we produced exhibit linear correlations. The correlation coefficients are shown in Table 3.1. All scatter plots are given in Appendix A.

Among 15 scatter plots, the two depicted in Figure 3.2 constitute the ones with the largest and the lowest correlation, respectively. Assuming a constant evolutionary rate within each orthologous family, the data points in the scatter plots are expected to be drawn from a straight line passing the origin of ordinates. Still, each data point is made up of two distance estimates each of which is subject to a measurement error.

We fit a regression line through the origin to the data points in a scatter plot. Since there is no clear assignment of dependent and independent variables, we perform a total least squares regression. Minimizing the sum of squared perpendicular distances of data points to a regression line is achieved by singular value decomposition. The variances of the two distance estimates that make up a data point are computed using the inverse Fisher information as described in [Lindgren, 1993] and [Müller, 2001; Müller *et al.*, 2002]. We choose the larger variance and assign it to a data point as isotropic variance in the scatter plot. To ensure that data points adequately contribute to the regression, the data points are scaled to unit variance. That is, the distances are divided by the square root of the isotropic variance prior to the regression [Press *et al.*, 1999].

We observe that the deviations of the majority of data points in the HC-FC plot from the regression line are in the range of the standard deviations (obtained as the square roots of the isotropic variances). The distances of 2283 (of 3640) data points to the regression line are smaller than the respective standard deviations. In contrast, there are only 708 data points in the HF-DC plot with a distance to the regression line that is smaller than the respective standard deviation.

Consider the trees below the scatter plots that show the evolutionary paths being compared. While the HF-DC comparison does not include a common edge at all, the HC-FC comparison comprises a large time of common evolutionary history. The

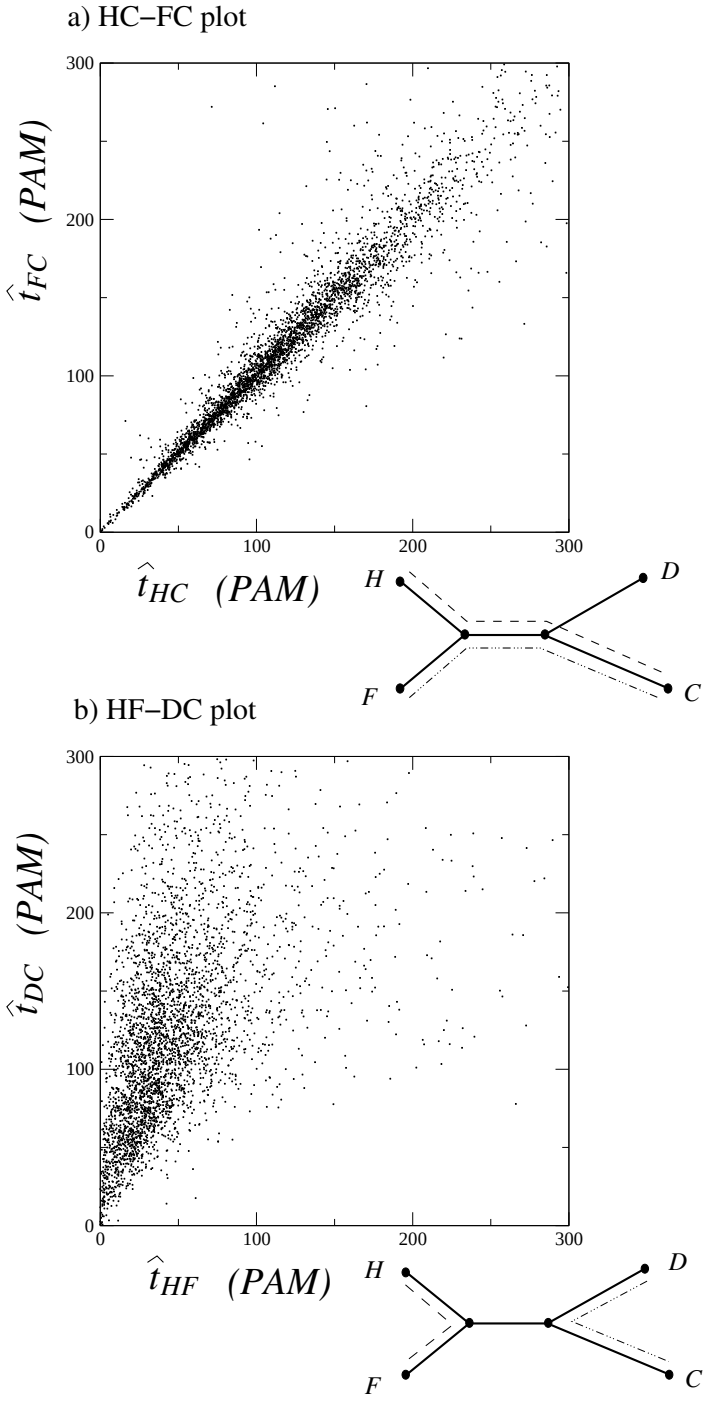

Figure 3.2: Scatter plots comparing evolutionary distances between proteins of orthologous families. A data point in the *HC-FC* scatter plot corresponds to the evolutionary distance $\hat{t}^i_{HC}$ between proteins of *H. sapiens* and *C. elegans* and to the distance $\hat{t}^i_{FC}$ between proteins of *F. rubripes* and *C. elegans* in orthologous family i. Dashed and dash-dotted lines in the trees shown below visualize the evolutionary paths compared in the scatter plots.

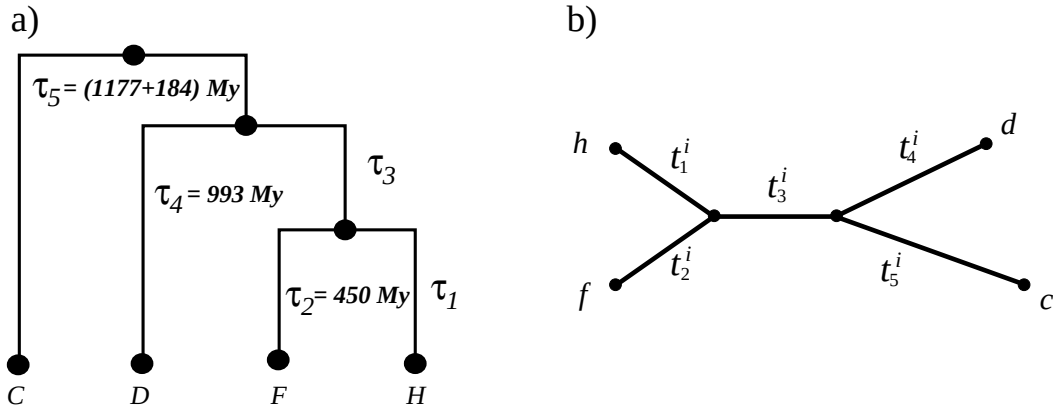

Figure 3.3: Species tree (a) and a gene tree (b) for the species under study. Edge lengths of the species tree are estimates of divergence times in Millions of years (My) [Wang *et al.*, 1999; Hedges, 2002]. The ultrametricity implies $\tau_1 = \tau_2$ and $\tau_3 = \tau_4 - \tau_2$.

larger dispersion of data points in the *HF-DC* plot is likely caused by lineage specific rate variations.

Linear correlations in the scatter plots, in particular in those that compare diverse evolutionary paths, empirically document that a fast or slow evolution is very much an attribute of a protein family.

3.5 Estimating family specific evolutionary rates

3.5.1 The tree length ratio

The literature provides estimates of divergence times for the species under study here [Wang *et al.*, 1999; Hedges, 2002]. Figure 3.3a shows the species tree with edge lengths representing times. We set up two estimators to measure family specific rates and use the divergence times to relate evolutionary distance measures to historical time.

Let Q be the rate matrix of the Müller-Vingron model, $\mathcal{X}_i$ the multiple alignment of orthologous family i, T the tree topology and let t^i_j , $j = 1, ..., 5$ denote the edge length parameters in the traditional likelihood computation for a phylogenetic tree (see Figure 3.3b).

First we do not make the assumption of rate constancy among lineages and no constraints are imposed on the edges of the phylogenetic tree. The edge length estimates $\hat{t}^i_j$ are the values where the likelihood function of equation 2.7 assumes its maximum. The tree length $\hat{t}^i = \sum_j \hat{t}^i_j$ of the maximum likelihood tree holds the total amount of substitutions having accumulated on the evolutionary paths. The time that has

passed since mutations accumulated is given by the tree length of the species tree $\tau = \sum_j \tau_j$. A natural estimator $\hat{l}_i$ for a family-specific evolutionary rate is given by the tree length ratio

$$\hat{l}_i = \frac{\hat{t}^i}{\tau} = \frac{\sum_j \hat{t}^i_j}{\sum_j \tau_j}. \tag{3.1}$$

3.5.2 The Family Specific Rate (FSR)

At the other extreme, we assume that the sequences have evolved according to the species tree and at one rate being constant over time and lineages. The parametrization

$$t^i_j = \lambda_i \cdot \tau_j, \qquad j = 1, .., 5. \tag{3.2}$$

yields the likelihood function

$$\mathcal{L}_{FSR}(\lambda_i) = \Pr(\mathcal{X}_i \mid \lambda_i, \tau_1, ..., \tau_5, T, Q). \tag{3.3}$$

The times $\tau_1, ..., \tau_5$ are fixed and λ_i scales the edges of the phylogenetic tree. It was previously suggested to apply the parametrization to smaller data sets [Yang, 1996]. We call the scaling factor $\hat{\lambda}_i$ that maximizes $\mathcal{L}_{FSR}$ the Family Specific Rate (FSR) of orthologous family i. Figure 3.4 illustrates the FSR estimation and the likelihood curvature for the cullin family.

Evolutionary distances t^i_j are measured in PAM units. We give the estimates of $\hat{l}_i$ and $\hat{\lambda}_i$ in units of PAM per billions of years (PAM/BYr).

A program to estimate Family Specific Rates was implemented in C. Optionally, the rates of bootstrap replicates on the multiple alignment are computed and the 95% bootstrap-confidence interval [Efron and Tibshirani, 1993] of $\hat{\lambda}_i$ is put out.

3.5.3 The Likelihood Ratio

The likelihood function $\mathcal{L}_{FSR}(\lambda)$ is recovered from the likelihood function $\mathcal{L}(t_1, ..., t_5)$ for the ML tree estimation when the parametrization of equation 3.5.2 is used. This indicates that the models to estimate $\hat{\lambda}_i$ and $\hat{l}_i$ are nested models. A likelihood ratio test (LRT) checks whether the simpler model assuming one constant rate of evolution is preferable.

The simpler model representing the null-hypothesis is the FSR-model $\mathcal{L}_0 = \mathcal{L}_{FSR}(\lambda_i)$ assuming that the sequences have evolved at one constant rate and according to the reference tree. The alternative model imposes no constraints on the rates and depends

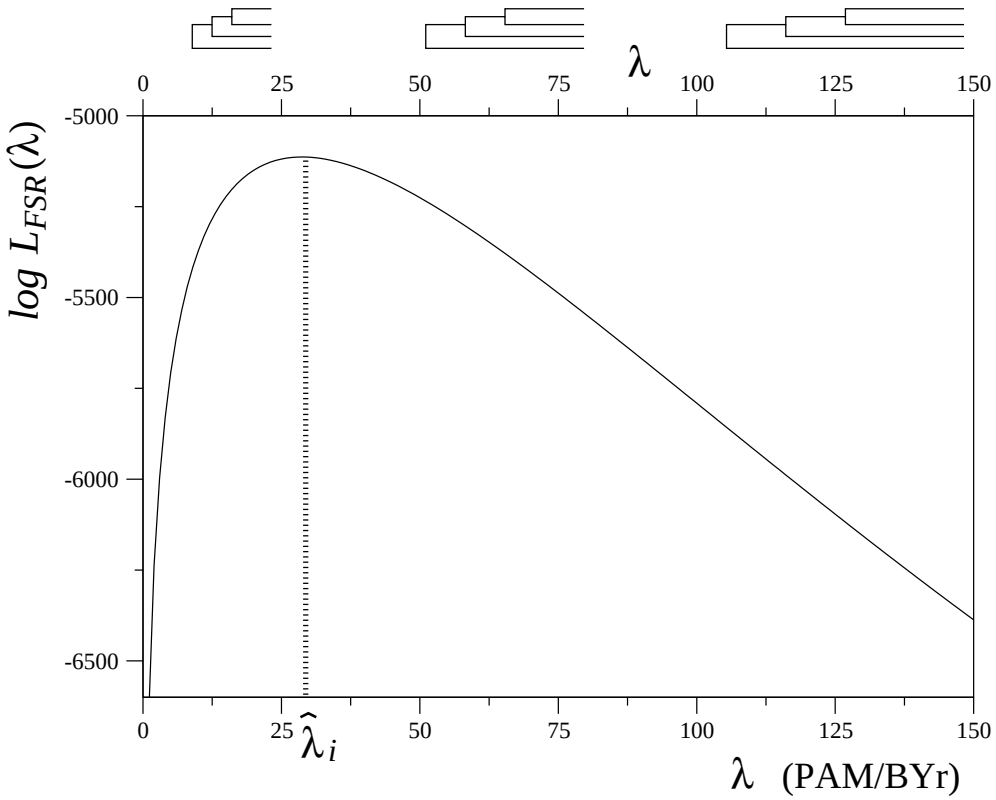

Figure 3.4: The log likelihood function $\log \mathcal{L}_{FSR}(\lambda)$ for the cullin family. The units of λ are PAM per billions of years (PAM/BYr). Schematic dendrograms at the top illustrate that λ scales edge lengths. The likelihood function assumes its maximum at $\hat{\lambda}_i = 29$ PAM/BYr. The 95% bootstrap-confidence interval ranges from 27 to 31 PAM/BYr.

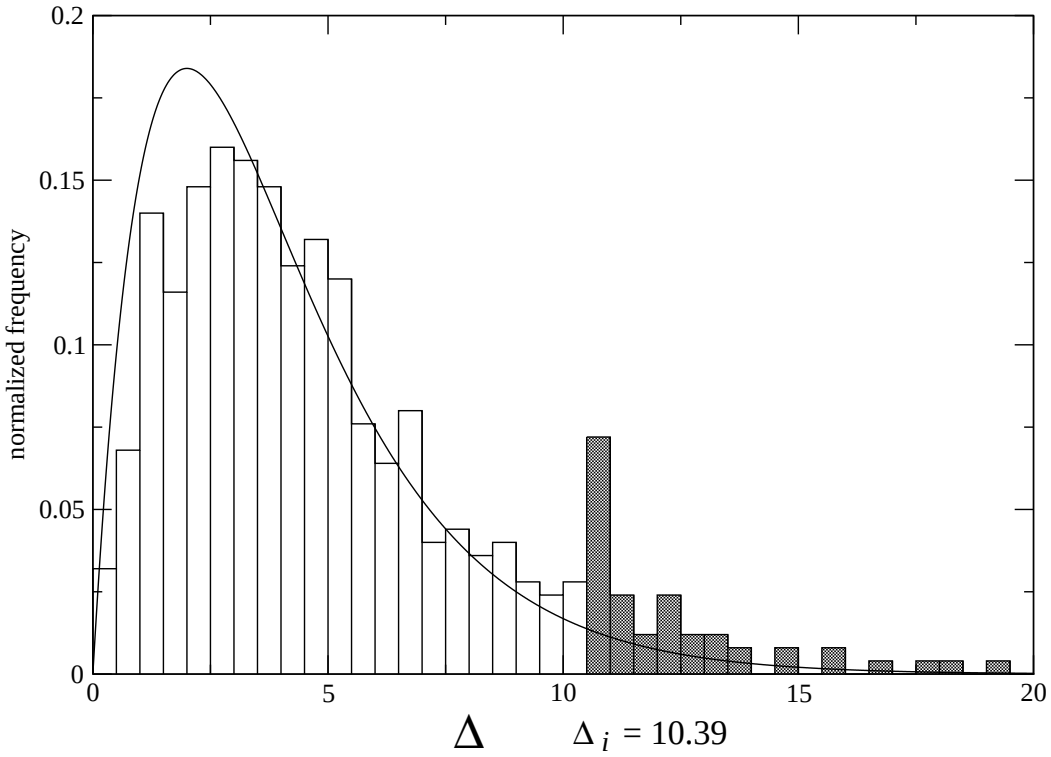

Figure 3.5: Simulated null distribution of log likelihood ratios Δ for an orthologous family of DNA replication helicases. Grey bars in the histogram correspond to Δ-values being larger than the original Δ_i. The simulated distribution is close to the the χ^2 distribution for four degrees of freedom (continuous line).

on five model parameters. Its likelihood function $\mathcal{L}_1$ is the likelihood of the ML tree $\mathcal{L}_1 = \mathcal{L}(t_1, .., t_5)$ when the topology T is fixed to the topology of the reference tree. The likelihood ratio statistic for family i is the difference $\Delta_i = 2(\log \mathcal{L}_1(\hat{t}_1^i, \ldots, \hat{t}_5^i) - \log \mathcal{L}_0(\hat{\lambda}_i))$ of the respective maximum log likelihoods. The difference in the number of parameters of the two models is 4. Thus, under the null-hypothesis the distribution of Δ_i is expected to be approximately χ^2 with 4 degrees of freedom, here denoted $\chi^2(4)$.

Since different families possess different characteristics, we do not assume that the distribution of Δ under the null-hypothesis is the same for all families. Instead of applying the χ^2 statistic we make the different Δ_i-values comparable across families by simulating family-specific null-distributions for Δ_i. For family i we do this by repeatedly (500 times) choosing one sequence from the family at random and using the reference tree and the rate matrix to simulate the rest of the family using Rose [Stoye *et al.*, 1998]. For each set of simulated families we then calculate the corresponding Δ-values and obtain a simulated p_i-value from the fraction of Δ-values being larger than Δ_i. We reject the null-hypothesis if $p_i < 0.05$.

Figure 3.5 shows the normalized histogram of 500 simulated Δ-values computed for an orthologous family containing DNA replication helicases. The log likelihood ratio computed from the alignment of the family is $\Delta_i = 10.39$. The grey bars correspond to 51 of 500 Δ-values being larger than Δ_i. We obtain $p_i = \frac{51}{500} = 0.102$ and accept the null-hypothesis. The continuous line in Figure 3.5 renders the $\chi^2(4)$ distribution. While $\chi^2(4)$ is close to the simulated distribution, its tail has a lower mass. The probability to observe $\Delta_i \geq 10.39$ under the $\chi^2(4)$ distribution is $\Pr(\Delta_i \geq 10.39) = 0.034$.

3.6 Results

3.6.1 FSR versus tree length

We estimate Family Specific Rates $\hat{\lambda}_i$ and tree length ratios $\hat{l}_i$ (by using TREE-PUZZLE [Strimmer and von Haeseler, 1996; Schmidt *et al.*, 2002]) and perform the likelihood ratio test for each orthologous family. The LRT reveals 888 orthologous families with $p_i \geq 0.05$. We call these families *rate constant families*.

The scatter plot in Figure 3.6 compares values of $\hat{\lambda}_i$ and $\hat{l}_i$ of all orthologous families. Interestingly both tree models yield almost the same rate estimates. As expected $\hat{\lambda}_i$ and $\hat{l}_i$ closely scatter around the bisecting line and assume virtually the same values for the rate constant families. Still values of $\hat{\lambda}_i$ and $\hat{l}_i$ for the whole set of families are also highly correlated with a correlation coefficient of $r = 0.982$. The different rate estimates are even close in 218 cases where a different topology than the one of the reference tree yields a higher likelihood in the ML tree computation. We conclude that rates of protein evolution are mainly driven by family specific effects. In the sequel the rate of an orthologous family is referred to by its Family Specific Rate $\hat{\lambda}_i$.

3.6.2 Rate distribution

On the basis of few mammalian sequences Kimura [1968] stated: "Averaging those figures for haemoglobin, cytochrome c and triosephosphate dehydrogenase gives an evolutionary rate of approximately one substitution in 28×10^6 yr for a polypeptide chain consisting of 100 amino acids[1]." His estimate fits well to the rates we derived. Figure 3.7 shows the overall distribution of Family Specific Rates $\hat{\lambda}_i$. The mean rate amounts to 52 PAM/BYr, the median rate to 50 PAM/BYr.

[1] 1 PAM / $28 \cdot 10^6 years \approx 36$ PAM per billions of years

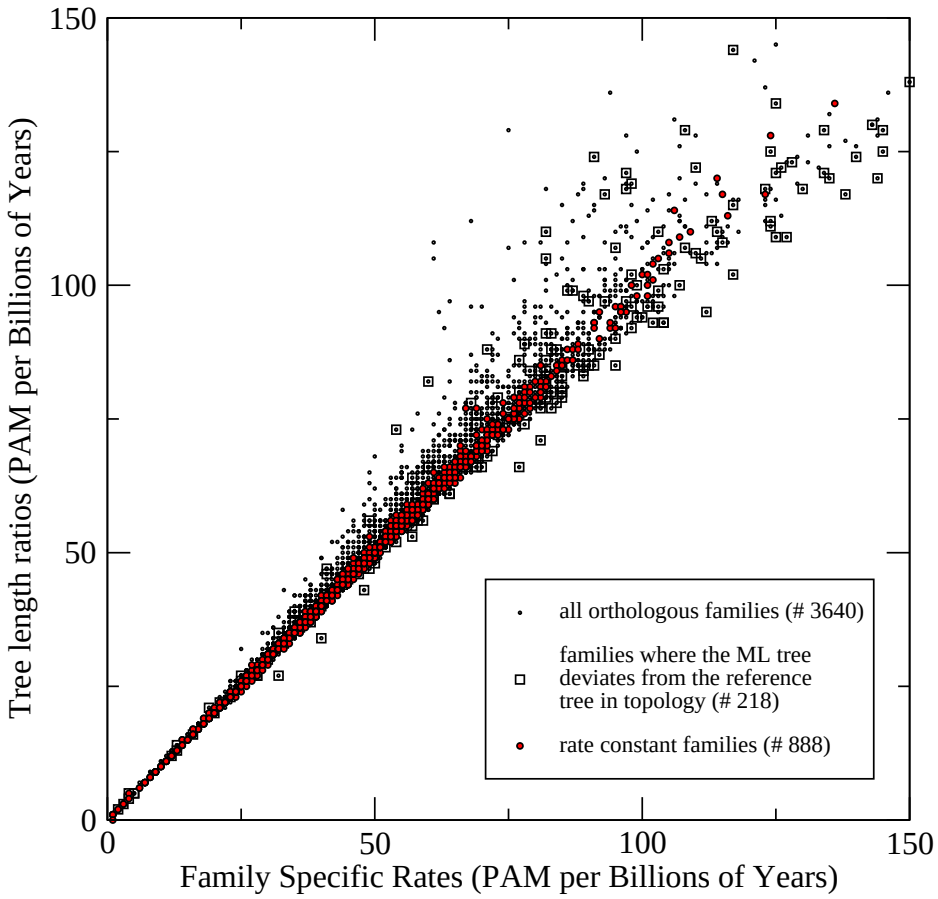

Figure 3.6: The scatter plot compares Family Specific Rates $\hat{\lambda}_i$ to tree length ratios $\hat{l}_i$.

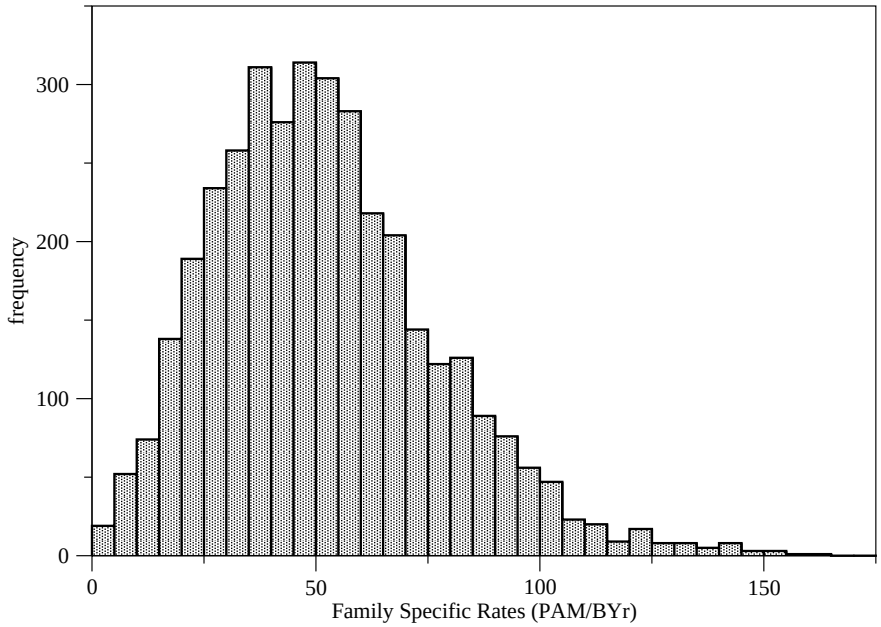

Figure 3.7: Histogram of all Family Specific Rates $\hat{\lambda}_i$.

Table 3.2 shows some examples of proteins with high and low ranking FSRs. The whole set of orthologous families, together with alignments and ML trees, FSRs, tree length ratios and p_i-values, is available available at `http://speeds.molgen.mpg.de`.

3.6.3 Family Specific Rates and nonsynonymous nucleotide substitutions

We compare Family Specific Rates to numbers of synonymous and nonsynonymous nucleotide substitutions. The two soil nematodes *Caenorhabditis briggsae* and *Caenorhabditis elegans* are closely related and diverged about 100 million years ago. The genome sequence of *C. briggsae* was completed in 2003 and exhaustively compared to the sequence of *C. elegans* [Stein *et al.*, 2003]. Stein *et al.* [2003] identified $12,155$ orthologous gene pairs using the reciprocal best blast hit and conserved synteny. The same authors aligned the orthologs and computed numbers of synonymous and nonsynonymous substitutions d_N and d_S using the codon substitution model. We adopt their results. They reason that cases where $d_N \geq 4$ or $d_S \geq 9$ are due to spurious alignments. We therefore accept their ML estimates for 9307 orthologous gene pairs. The intersection of orthologous gene pairs and *C. elegans* translated sequences being present in alignments of our data set amounts to 2320. Of those 2320 orthologous families, 592 have evolved at a constant rate according to the LRT.

SwissProt	**slowly evolving proteins**	FSR
P02307	Histone H4	1
P07181	Calmodulin	1
P02570	Actin	1
P40946	ADP–ribosylation factor 3	3
P35129	Ubiquitin–protein ligase	5
P48601	26S protease regulatory subunit 4	5
P06749	RAS–like GTP–binding protein RhoA	6
P48462	Serine/threonine protein phosphatase beta isoform	6
P17080	GTP–binding nuclear protein RAN	6
Q19877	40S ribosomal protein S23	7
	fast evolving proteins	
P14060	Trophoblast antigen FDO161G	104
Q08379	Golgi autoantigen	105
Q04637	Eukaryotic translation initiation factor 4 gamma	105
Q99853	Forkhead box protein B1 (Transcription factor FKH-5)	117
P22293	Suppressor of sable protein	125
P57682	Kruppel–like factor 3	126
Q99466	Neurogenic locus notch homolog protein 4	131
P82295	Prominin–like protein	135
P40197	Platelet glycoprotein V	145

Table 3.2: The table lists slowly and fast evolving proteins. Swiss-Prot [Gasteiger *et al.*, 2003] accession numbers on the left and the respective protein names refer to one of the aligned proteins of an orthologous family. Family Specific Rates on the right are estimated in PAM/BYr units.

Family Specific Rates $\hat{\lambda}_i$ and estimated nucleotide substituions are positively correlated (see Table 3.3). Nonsynonymous substitutions d_N cause a change in the amino acid sequence. The fact that d_N and $\hat{\lambda}_i$ are not perfectly correlated is primarily due to nematode-specific rate variations. The scatter plot in Figure 3.8 compares $\hat{\lambda}_i$ and d_N for the rate constant families. Here the correlation coefficient is larger.

We use those 592 rate constant families to test whether the rate measures agree in magnitude. For this purpose we relate the two measures and fit a regression line through the origin of ordinates to the data points of the scatter plot. Consider two data points that are located on the same straight line passing the point of origin. Both of them favor the same slope of the regression line. We therefore first project the data points to the unit circle to let them adequately contribute to the regression line. Second we perform a total least squares regression to minimize the sum of squared

	d_N	d_S	$\omega = \frac{d_N}{d_S}$
Family Specific Rate (FSR)	0.460	0.325	0.328
FSR (rate constant set)	0.526	0.359	0.394

Table 3.3: Correlation coefficients when comparing Family Specific Rates to numbers of nonsynonymous and synonymous substituions d_N and d_S between *C. elegans* and *C. briggsae*.

euclidian distances of data points to the regression line (see also Section 3.4).

The slope of the fitted regression line is

$$m = 1.75 \cdot 10^{-3} \ \text{BYr/PAM}$$

To check whether the magnitudes of the rate measures agree, we estimtate the divergence time of *C. elegans* and *C. briggsae* using the slope of the regression line. We transform d_N^i, the number of nonsynonymous substitutions mapped to orthologous family i, into a rate r_i with unit PAM/BYr. Since d_N^i is measured on orthologous sequences from *C. elegans* and *C. briggsae* we divide d_N^i by two times the divergence time of the two nematodes τ_n (in billions of years). This ratio is multiplied with 100 PAM as d_N^i holds the number of nonsynonymous substitutions per one nonsynonymous site only:

$$r_i = \frac{d_N^i}{2 \cdot \tau_n} \cdot 100 \text{ PAM}$$

Consider that $\hat{\lambda}_i$ and r_i for orthologous family i assume the same value and that the data point $(\hat{\lambda}_i, d_N^i)$ is drawn from the regression line. We obtain

$$m = \frac{d_N^i}{\hat{\lambda}_i} = \frac{2 \cdot r_i \cdot \tau_n}{\hat{\lambda}_i \cdot 100 \text{ PAM}} = \frac{2 \cdot \tau_n}{100 \text{ PAM}} = 1.75 \cdot 10^{-3} \ \frac{\text{BYr}}{\text{PAM}} \ .$$

Solving for the divergence time τ_n of *C. elegans* and *C. briggsae* yields $\tau_n = 87.5$ millions of years. Our estimate is in the range of previous estimates. For example Coghlan and Wolfe [2002] estimate a divergence time around 50–120 millions of years. In this particular case the Family Specific Rates obtained from aligned proteins and the measures of nonsynonymous nucleotide substitutions agree in magnitude.

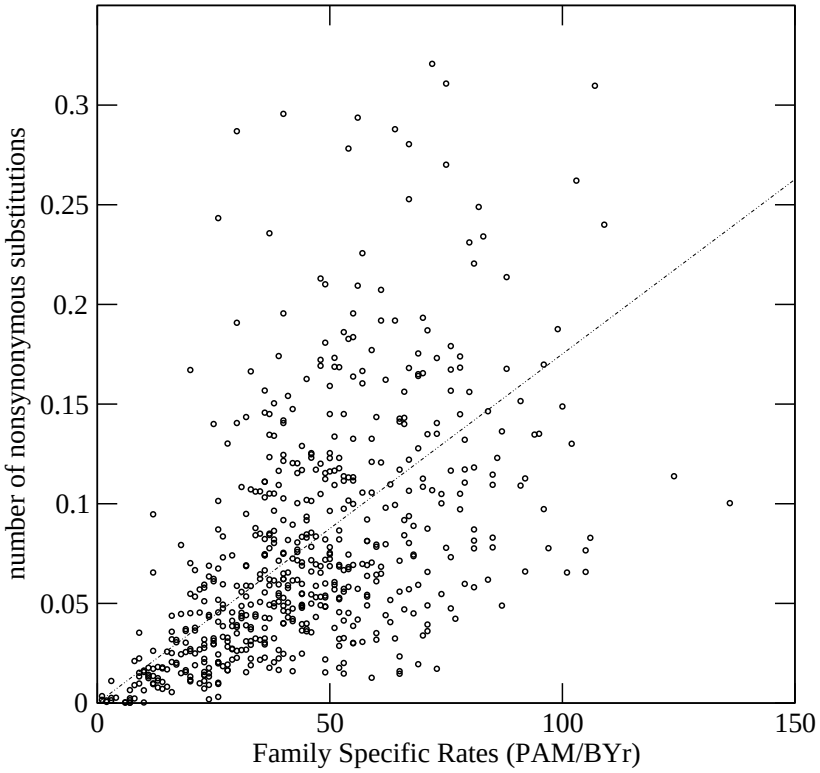

Figure 3.8: Scatter plot comparing Family Specific Rates of rate constant families to numbers of nonsynonymous substitutions d_N between *C. elegans*-*C. briggsae* orthologs. The slope of the regression line is used to test whether the rate measures agree in magnitude.

.

phenotype class	total number	number of FS rates	mean FSR (PAM/BYr)	ranksum p-value
All	–	3640	52.4	–
Nonv	1170	502	39.8	$1.29 \cdot 10^{-28}$
Grow	276	117	52.4	0.902
Vpep	276	68	47.9	0.115

Table 3.4: Mean Family Specific Rates for orthologous families when they are grouped according to one of the three phenotype classes observed for the *C. elegans* sequence. The p-value in the rightmost column is obtained by comparing the rates of genes within a phenotype class to all rates by a Wilcoxon two sample test.

3.6.4 Rates of essential genes, RNA interference

Essential genes can be spotted by knock-out experiments. If the absence of a gene results in a lethal or sterile phenotype the gene is considered essential. Such genes are expected to be subject to stringent purifying selection [Cutter *et al.*, 2003].

Small double-stranded RNA molecules can interfere the translation of mRNA molecules obeying a similar sequence. The small RNA molecules are called short interfering RNAs (siRNAs) and the mechanism is known as RNA interference (RNAi). A 'genome-wide' loss-of-function analysis covered 86% of *C. elegans* genes [Kamath *et al.*, 2003]. According to the observed phenotype the genes were grouped into three classes: "the nonviable class (Nonv), consisting of embryonic or larval lethality or sterility (with or without associated post-embryonic defects); the growth defects (Gro) class, consisting of slow or arrested post-embryonic growth; and the viable post-embryonic phenotype (Vpep) class, consisting of defects in post-embryonic development (for example, in movement or body shape) without any associated lethality or slowed growth" [Kamath *et al.*, 2003].

We compare Family Specific Rate distributions being specific to a phenotypic class observed for the *C. elegans*-sequence in the alignment of the orthologous family. Table 3.4 summarizes the results. The nonviable class is the only phenotype class where significant differences in rate distributions are observed. Of 1170 worm genes within the nonviable class, 502 genes are found within our alignments of orthologous families. The rate distributions are shown in Figure 3.9. Most of the nonviable genes are subject to stringent purifying selection. The mean rate of the *C. elegans*-nonviable set amounts to 39.8 PAM/BYr. The Wilcoxon two sample test comparing the overall to the nonviable rate distribution yields a p-value of $p = 1.29 \cdot 10^{-28}$ (see Table 3.4).

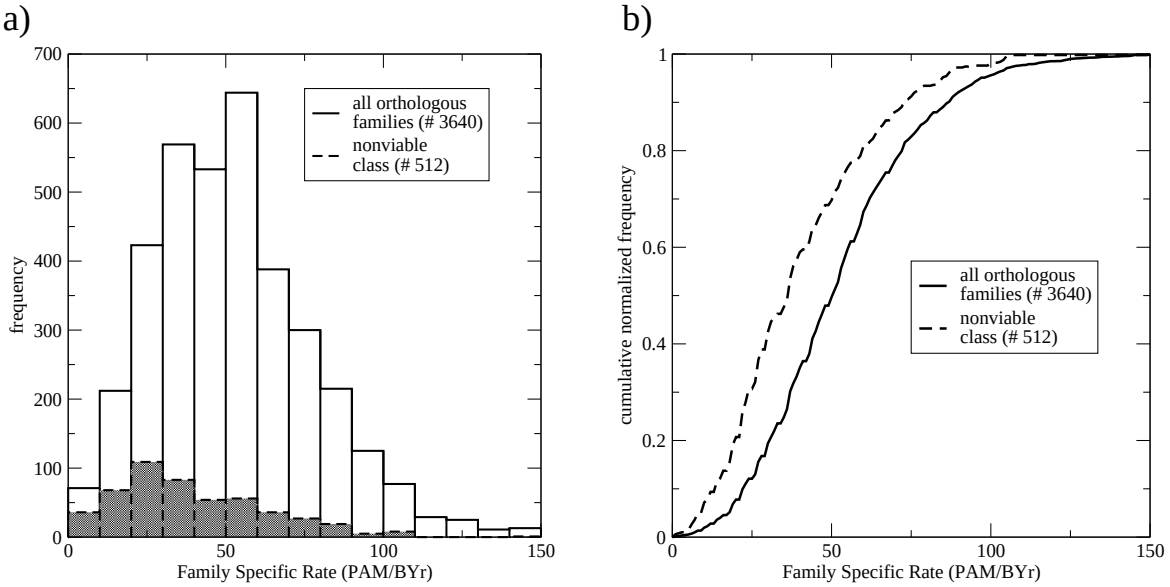

Figure 3.9: a) The overall rate distribution compared to the rate distribution of orthologous families where the sequence from *C. elegans* had a lethal or sterile phenotype. b) The same distributions as cumulative normalized histograms.

3.6.5 Protein interaction and the rate of evolution

Thousands of different proteins are active in a cell at any time. Some of them act as single monomeric units. Most of the proteins are only functional as groups. Some proteins are components of large complexes. Others function in association with partner molecules, e.g., by transporting signals from one cellular component to another or by occupying a specific compartment and receiving the signals. Proteins mutually affect their functions by interaction. Networks that represent those interactions are called *interactome maps*.

In 2004, a large fraction of the *C. elegans* interactome was reliably mapped [Li *et al.*, 2004]. The authors obtained more than 4000 interactions through carefully performed high throughput two-hybrid analysis. Already known and further potential interactions predicted from orthologs of other organisms were added and alltogether 5534 interactions for 2898 proteins were combined into the Worm Interactome version 5 (WI5). Like other biological networks, the worm interactome exhibits scale-free properties. A putative connection between the scale-free topology and genetic robustness was reported for the yeast proteome: As a consequence of knocking out genes representing hubs, the mutant phenotype is likely lethal [Jeong *et al.*, 2001; Wuchty, 2002; Han *et al.*, 2004]. Correlations of evolutionary rates to numbers of interaction partners were found to be weak and a putative relation is controversely discussed [Hurst and Smith, 1999; Jordan *et al.*, 2003; Bloom and Adami, 2003; Fraser and Hirsh, 2004;

	$k \in \{1\}$	$k \in \{2,3\}$
$k \in \{2,3\}$	0.003	
$k \in \{4,..,89\}$	$1.04 \cdot 10^{-4}$	0.35

Table 3.5: p-values of Wilcoxon two sample tests when comparing Family Specific Rates between three sets of orthologous families with different numbers of interaction partners k.

Bloom and Adami, 2004].

Do the number of interactions within WI5 correlate to evolutionary rates? In the following, the number of interaction partners is referred to as degree k. We find 765 of 2898 worm genes in WI5 in our data set. Rates and degrees are weakly negatively correlated. A relation is established when partitioning the set of 765 proteins with respect to the degree of the worm proteins and to the rates, respectively.

First we split those 765 proteins into three sets with degrees $k \in \{1\}$, $k \in \{2,3\}$ and $k \in \{4,..,89\}$ respectively and compare the rate distributions among the three sets. We find 380 proteins with degree $k \in \{1\}$, 199 with degree $k \in \{2,3\}$ and 186 with degree $k \in \{4,..,89\}$. Indeed, the average rate of the three sets decreases with growing k, suggesting that purifying selection acts stronger on hubs of the interactome. Table 3.5 lists the p-values of Wilcoxon two sample tests when comparing the rate distributions of the three sets. It turns out that the rate distributions of the sets for $k \in \{2,3\}$ and $k \in \{4,..,89\}$ do not significantly differ. Yet the comparison of both of them to the rate distribution of families with $k = 1$ yields a significant p-value.

Second we split the set of 765 orthologous families into four approximately same sized sets with rates in four different non-intersecting rate intervals. The bar chart in Figure 3.10 compares the frequencies of families for a given rate interval and a certain degree category. For $k \in \{1\}$ we observe that most of the families belong to the fastest rate interval. For $k \in \{4,..,89\}$ the reverse holds.

Our results support the view that interactions impose additional constraints on the replacement of amino acid residues.

3.6.6 Does protein function constrain the rate?

Rates of transcription factors and enzymes

In order to detect and to eventually further explore a putative connection of Family Specific Rates to specific protein functions, we investigate rate distributions of enzymes, transcription factors and proteins carrying disulfides. The three sets of orthologous families were obtained by exploiting Swiss-Prot annotations [Gasteiger *et al.*,

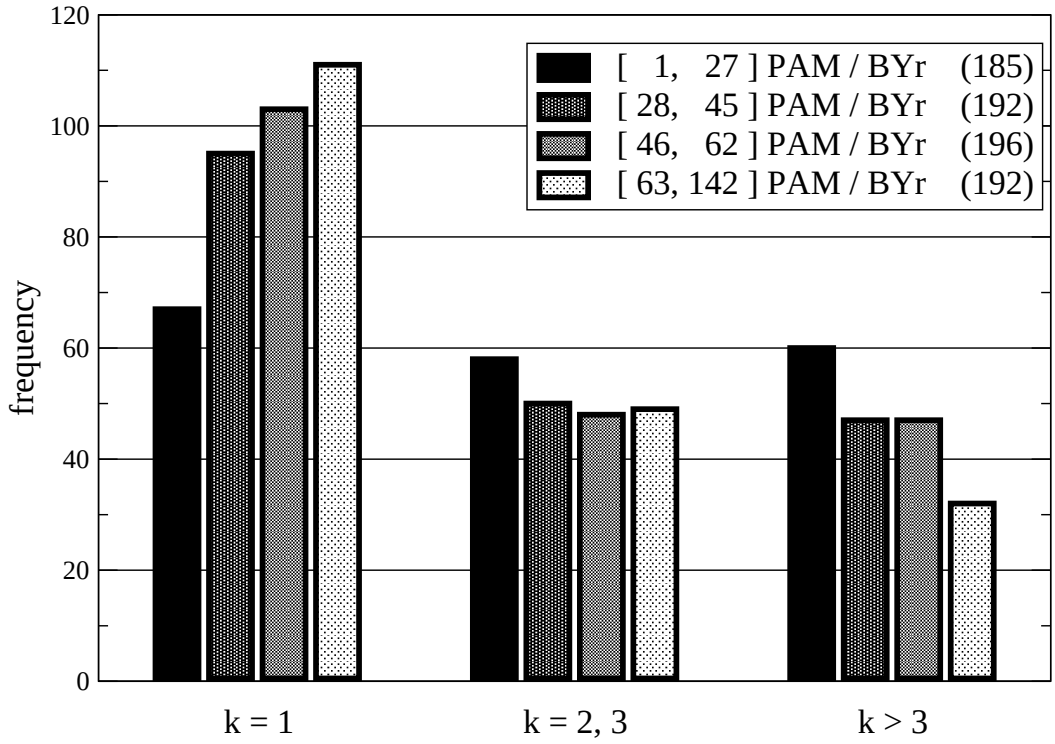

Figure 3.10: The bar chart compares Family Specific rates to numbers of interaction partners in the WI5 data set. 765 orthologous families were defined as belonging to one of four rate categories as well as to one of three degree categories. Numbers in parentheses indicate the numbers of orthologous families belonging to a rate category. For degree $k \in \{1\}$, most of the orthologous families are fast evolving. For proteins with $k \in \{4, .., 89\}$ small rates are over-represented.

2003]. Swiss-Prot is a protein sequence database with high-quality curated annotations. Since we frequently want to use Swiss-Prot provided information we map 5807 Ensembl-sequences of our data set to accessions of the Swiss-Prot Release 43.0 by requiring that the sequences shared at least 98% identical residues over 90% of their length.

Table 3.6 summarizes some values of the resulting rate distributions. We identify 164 families where the sequences could be linked to a transcription factor in TRANSFAC [Matys *et al.*, 2003]. Castillo-Davis *et al.* [2004] point to the possibility that transcription factors are overrespresented among the fastest evolving proteins. In accord with results presented of Koonin *et al.* [2004], we find that the rate distribution of transcription factors does not deviate from the overall rate distribution.

The situation is different for the set of 811 families with sequences being referred to in the ENZYME database [Bairoch, 2000]. The mean rate of enzymes is clearly below the mean rate of all families. A Wilcoxon two sample test with the null hypothesis that the enzyme rate distribution and the overall rate distribution are drawn from the same sample yields $p = 9.24 \cdot 10^{-24}$. We will return to the enzymes and their rates in Section 4.4.6.

Proteins with disulfide bridges are fast evolving

We identify 112 orthologous families with annotated *disulfide bridges* (or *disulfide bonds*) in Swiss-Prot. Disulfide bridges endow the proteins with greater stability. A covalent bond between the sulphur atoms of cysteines is formed during protein folding when the peptide is exposed to an oxidative milieu, e.g., in the lumen of the rough ER. The rate distribution of proteins with disulfide bridges is significantly shifted to large rates. The fact that disulfide bridges constitute the dominating structural element of extra-cellular proteins and are rarely found under reducing intracellular conditions, motivates to inspect the rate distributions of proteins when they are grouped according to their subcellular localization.

Rates according to subcellular localization

We search the sequences of our data set for SMART domains by using "hmmsearch" from the HMMER package and considering hits with E-values below the SMART provided E-value for the lowest scoring true positive. Subsequently we assign the orthologous families to the following subcellular locales: "nuclear", "cytoplasmic" or "secreted" if exclusively nuclear, signalling or extra-cellular domains were present in the family; "nuclear_cytoplasmic" or "cytoplasmic_secreted" if either nuclear and signalling or signalling and extra-cellular domains were present. From Table 3.6 we see

subset	subset size	mean rate	standard deviation	median rate	ranksum p-value
all orthologous families	3640	52.4	25.1	50	-
transcription factors	164	54.4	25.2	49	0.478
enzymes	811	43.2	20.3	40	$9.42 \cdot 10^{-24}$
disulfide bridges	112	64.7	24.6	62	$1.93 \cdot 10^{-7}$
nuclear	563	52.9	27.0	49	0.844
cytoplasmic	601	50.3	25.4	48	0.066
nuclear_cytoplasmic	108	58.8	25.3	53	0.014
secreted	90	68.5	25.8	66	$5.24 \cdot 10^{-9}$
cytoplasmic_secreted	17	74.8	25.0	77	$3.25 \cdot 10^{-4}$

Table 3.6: The table lists mean rates, standard deviations and median rates of subsets of all orthologous families. Units are PAM/BYr. The p-value in the rightmost column is obtained by comparing the rates of a subset to all rates by a Wilcoxon two sample test.

that neither a putative nuclear nor a cytoplasmic locale significantly changes evolutionary rates. Only the rates of the "secreted" and the "cytoplasmic_secreted" sets are significantly different when compared to rates of all orthologous families.

Winter *et al.* [2004] and Koonin *et al.* [2004] confirm the observation that secreted proteins are subject to a rapid accumulation of mutations. We focus on extra-cellular proteins in greater detail and set up a working hypothesis being named like the caption of the next chapter.