Family specific rates of protein evolution

Hannes Luz and Martin Vingron

Department of Computational Molecular Biology
Max-Planck-Institut für molekulare Genetik, Ihnestr. 73, 14195 Berlin, Germany
{Hannes.Luz|Martin.Vingron}@molgen.mpg.de

Abstract:
Amino acid changing mutations in proteins are contstrained by purifying selection and accumulate at different rates. We estimate evolutionary rates on multiple alignments of eukaryotic protein families in a maximum likelihood framework. We find that the evolution of indispensable proteins is constrained by selection and that protein secretion is coupled to an increased evolutionary rate.

1 Introduction

Proteins evolve at different rates because the mutational process that acts on a protein is subject to a specific selective pressure. For example, an alignment of histones from man, fugu, fly, and worm shows only few amino acid exchanges. Here almost all amino acid changing substitutions are deleterious, the selective regime is rather stringent. At the other extreme orthologous receptor tyrosine kinases from the same organisms may be even non-trivial to align and only few residues are under strong purifying selection.

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 [Yan97] to estimate d_N, the expected number of non-synonymous substitutions causing a change of the amino acid sequence [DP04, CDKHK04, HK05]. 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 measures of d_N for individual orthologous gene pairs constitute a 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 [DP04], the 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 larger sets of orthologous proteins, that is, within *orthologous families*. Koonin *et al.* [KFJ$^+$04] apply a measure for an evolutionary rate on sets of distantly related orthologs by averaging distances from the outgroup sequence to other sequences. In our study, evolutionary rates of orthologous families are estimated in a maximum likelihood (ML) framework. We require the orthologous

families to hold members of a defined set of organisms. Thus, the total time that has passed since the sequences diverged is the same within different orthologous families and different levels of sequence divergence can be compared and related to historical time.

The promising goals when pinpointing rate distributions include the disclosure of global principles influencing selection. For example, purifying selection is expected to act weaker on dispensable than on indispensable proteins and some authors accomplish correlating some degree of a protein's dispensability to its evolutionary rate [HF01, HK05]. Others try to relate evolutionary rates to sequence length, tissue specificity, secretion or the affiliation of the proteins to functional categories [LSK$^+$02, WGP04, KFJ$^+$04, CDKHK04].

This paper is structured as follows. Section 2 outlines the acquisition of the orthologous families and multiple alignments. Section 3 presents two ML estimators for a family specific rate. In Section 4, the prevalent assumption that indispensable proteins are more evolutionary conserved is put forward. Finally, we investigate the rates of extra-cellular proteins in Section 5. A case study focuses on the rates of protein tyrosine kinases.

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.

The peptide sequences were downloaded from the *Ensembl* database (version 16) [HBB$^+$02]. We first apply the INPARANOID software to obtain orthologous groups for each pair of organisms by requiring a high confidence for orthologous assignments and setting the INPARANOID confidence value to 95% [RSS01]. 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.

We select the orthologous families that contain at least one representative of each organism. When an orthologous family contains more than one sequence per organism we select four sequences of similar length. Sequences are filtered for low complexity regions [WF93] and for each family a multiple alignment of four orthologous sequences is generated using DCA [SMD97]. 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. 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. We end up with a set of 3640 orthologous families and multiple alignments. For the ML tree computations described below we consider the gapless sites of the alignments only.

3 Estimating family specific rates of protein evolution

We apply two approaches to estimate family specific evolutionary rates using standard ML phylogenetic tree estimation procedures. Under the assumption that point mutations accumulate according to a stochastic process that acts independently on the sites of a sequence, a reversible Markov process with a stationary distribution is commonly used as a probabilistic model of sequence evolution [MV00, MSV02]. We choose the Müller–Vingron model as amino acid replacement model where replacement frequencies were estimated on alignments of varying degree of divergence [MSV02]. Further, the Markov process is calibrated to PAM units. In concrete terms, in a sequence of 100 residues that evolves according to the Markov process and along an edge in a model tree with length $t = 1$ PAM one substitution event is expected to occur.

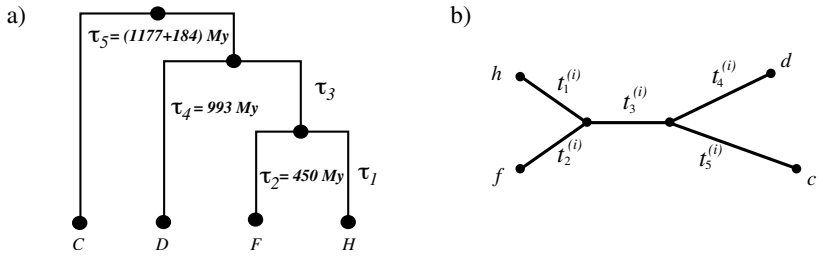

Figure 1: Species tree (a) and a gene tree (b) for the species under study. Edge lengths of the species tree $\tau_j,\ j = 1, ..., 5$ are estimates of divergence times in Millions of years (My) [WKH99, Hed02]. Species names are abbreviated by C (*Caenorhabditis elegans*), D (*Drosophila melanogaster*), F (*Fugu rubripes*), and H (*Homo sapiens*). Gene names of family i are abbreviated by c, d, f, and h. The ultrametricity implies $\tau_1 = \tau_2$ and $\tau_3 = \tau_4 - \tau_2$.

For the species under study here, the unrooted tree topology T of the species phylogeny is known. Consider the likelihood $\mathcal{L}(\vec{t}^{(i)})$ of the phylogentic tree [Fel81] for orthologous family i where four orthologous sequences are placed at the leaves of a tree (see Figure 1(b))

$$\mathcal{L}(\vec{t}^{(i)}) = \Pr(\mathcal{X}^{(i)} \mid \vec{t}^{(i)}, T, Q) = \prod_{s=1}^{n^{(i)}} \Pr(\mathcal{X}_s^{(i)} \mid \vec{t}^{(i)}, T, Q) \tag{1}$$

$\mathcal{L}(\vec{t}^{(i)})$ is the probability to observe the alignment $\mathcal{X}^{(i)}$ with the aligned positions $\mathcal{X}_s^{(i)}$, $s = 1, ..., n^{(i)}$, that have evolved according to the tree T with edge lengths $\vec{t}^{(i)} = (t_1^{(i)}, ..., t_5^{(i)})$ under our evolutionary Markov process Q. (In the following, we omit the notation of T and Q that remain fixed in our likelihood computations.)

The literature provides estimates of divergence times for the species under study here [WKH99, Hed02]. Figure 1(a) shows the species tree with edge lengths $\vec{\tau} = (\tau_1, ..., \tau_5)$ representing times. We use these divergence times to relate measures of amino acid replacements in a phylogenetic tree to historical times.

First, we do not assume rate constancy among lineages and no constraints are imposed on the edge lengths of the phylogenetic tree. The edge length estimates $t_j^{(i)} = \hat{t}_j^{(i)}$, $j = 1, ..., 5$, are the values where the likelihood function $\mathcal{L}(\vec{t}^{(i)})$ assumes its maximum. The tree length $\sum_j \hat{t}_j^{(i)}$ 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 $\sum_j \tau_j$. Thus, a natural measure for a family-specific evolutionary rate is given by the *tree length ratio* $\hat{l}_i$

$$\hat{l}_i = \frac{\sum_j \hat{t}_j^{(i)}}{\sum_j \tau_j} \tag{2}$$

At the other extreme, we assume that the sequences have evolved at a constant rate along the edges of the species tree in Figure 1(a). With the parametrization

$$\vec{t}^{(i)} = \lambda_i \cdot \vec{\tau} \tag{3}$$

(suggested, e.g., in [Yan96]) the likelihood depends just on the parameter λ_i

$$\mathcal{L}_\lambda(\lambda_i) = \prod_{s=1}^{n^{(i)}} \Pr(\mathcal{X}_s^{(i)} \mid \lambda_i \vec{\tau}) \tag{4}$$

We call the scaling factor $\hat{\lambda}_i$ that maximizes $\mathcal{L}_\lambda(\lambda_i)$ the *Family Specific Rate* (FSR) of family i.

How well does the Family Specific Rate model fit the data? This question can be adressed either for the individual orthologous families or for the whole data set that contains all families. In the latter case, we consider the alignments of all 3640 orthologous families in a large concatenated alignment $\mathcal{X}$. Competing model assumptions can be tested by likelihood ratio tests (LRT), the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) [Fel04]. Under the assumption that the sequences in $\mathcal{X}$ have evolved according to our species tree and at Family Specific Rates, the total maximum likelihood in the FSR model is $\mathcal{L}_{\text{FSR}} = \prod_{i=1}^{3640} \mathcal{L}_\lambda(\hat{\lambda}_i)$. When applying the above mentioned tests, the FSR-model performs well in comparison to other models. For example, we make use of PAML [Yan97] and assume a model where the rate parameter at an alignment position of $\mathcal{X}$ is drawn from a discretized gamma distribution with 40 rate categories. The latter model depends on the five edge lengths of the tree estimated and on the shape parameter of the gamma distribution. In a comparision of this model to the FSR model, the many more rate parameters in the FSR model are penalized by the AIC and by the BIC. Still, according to AIC and BIC the FSR model fits the data better.

For the individual families the likelihood function $\mathcal{L}_\lambda(\lambda_i)$ is recovered from the likelihood function $\mathcal{L}(\vec{t}^{(i)})$ with the parametrization in equation 3. This indicates that the models to estimate $\hat{\lambda}_i$ and $\hat{l}_i$ are nested models. Thus, a LRT that checks whether the simpler FSR model is preferable can be carried out. We estimate Family Specific Rates $\hat{\lambda}_i$ and tree length ratios $\hat{l}_i$ and perform a LRT for each orthologous family. The LRT reveals 888

orthologous families that have evolved at an approximately constant rate according to our species tree. We call these families *rate constant families*.

The scatter plot in Figure 2(a) 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 $r = 0.982$.

We conclude that the rates of protein evolution are driven by family specific effects. In the following we refer to the rate of an orthologous family by its Family Specific Rate $\hat{\lambda}_i$. Figure 2(b) shows the overall distribution of Family Specific Rates $\hat{\lambda}_i$ ranging from 1 to 162 PAM per billions of years (PAM/BYr). The mean rate amounts to 52 PAM/BYr, the median rate to 50 PAM/BYr.

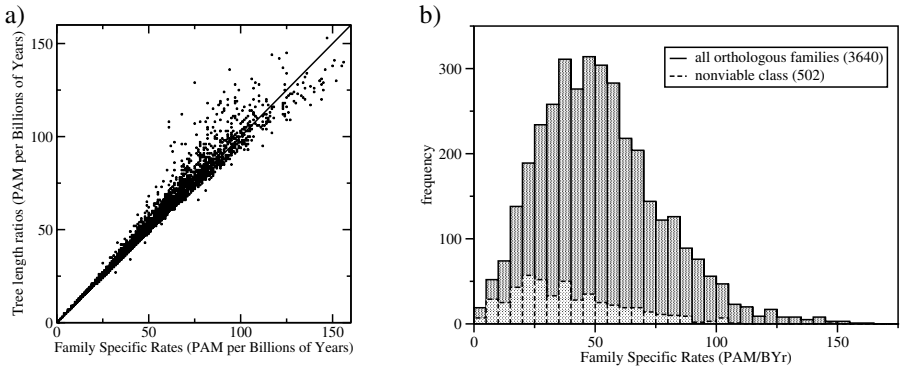

Figure 2: a) The scatter plot compares Family Specific Rates $\hat{\lambda}_i$ to tree length ratios $\hat{l}_i$. b) Histogram of all Family Specific Rates $\hat{\lambda}_i$ and of the subset of rates for families in the nonviable class (see Section 4).

4 Indispensable proteins are more conserved

Essentiality of a gene can be identified 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 [CPS+03]. 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 [KFD+03]. 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" [KFD$^+$03].

phenotype class	total number	number of FS rates	mean FSR (PAM/BYr)	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 1: Mean Family Specific Rates for orthologous families when they are grouped according to one of the three phenotype classes observed in 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.

We compare rate distributions of orthologous families with proteins in specific phenotypic classes to the overall rate distribution. Table 1 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 histograms in Figure 2(b) illustrate that the families of the nonviable class have lower rates compared to the rates of all orthologous families. 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 1).

Further, we investigate the interrelation of our evolutionary rates and a large fraction of the *C. elegans* interactome. Li *et al.* [LAB$^+$04] 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 WI5 exhibits scale-free properties.

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 with respect to the rates.

First we split the 765 proteins into three sets with degrees $k \in \{1\}$, $k \in \{2, 3\}$, and $k \in \{4, .., 89\}$ 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 mean rate of the three sets decreases with growing k, suggesting that purifying selection acts stronger on hubs of the interactome. 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 together or individually to the rate distribution of families with $k \in \{1\}$ yields significant p-values (e.g., for the sets with $k \in \{1\}$ and with $k \in \{4, .., 89\}$ the p-value is $p = 1.04 \cdot 10^{-4}$).

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 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.

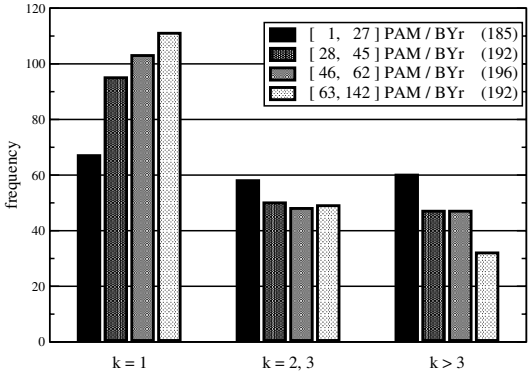

Figure 3: The bar chart compares Family Specific Rates to numbers of interaction partners in the WI5 data set. 765 orthologous families were defined as belonging to one of four rate categories as well as to one of three degree categories. Numbers in parentheses indicate the numbers of orthologous families 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.

5 Extra-cellular proteins are fast evolving

We combine different *in silico* approaches to derive a set of extra-cellular families. Namely we check putative extra-cellular localization due to Swiss-Prot annotations [NR02], the detection of extra-cellular SMART domains [MSBP02, LCS$^+$04] and the existence either of a predicted signal peptide [NEBvH97] or a transmembrane helix [SvHK98]. If the proteins of an orthologous family meet at least two of those criteria, we call the respective family *extra-cellular*.

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. We observe that the rate distribution of the extra-cellular families is significantly shifted to larger rates. The mean rate and the median rate are 67.2 PAM/BYr and 64 PAM/BYr, respectively. 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 $p = 1.80 \cdot 10^{-19}$.

Protein Tyrosine Kinases (PTKs) are involved in cellular signalling pathways and regulate key cell functions such as proliferation, cell growth, immune response and differentiation.

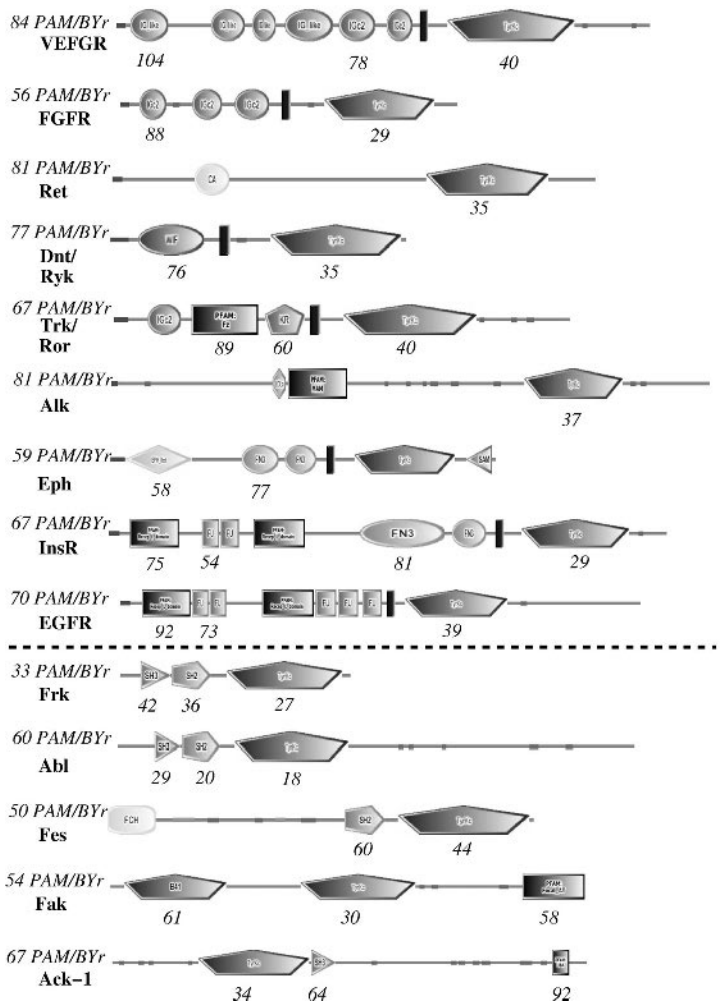

Figure 4: Receptor Tyrosine Kinases (above dashed line) and Protein Tyrosine Kinases (below dashed line). Representations of the proteins by their domain architectures were downloaded from the SMART web server [LCS+04] and comprise predicited SMART domains (bubbles), PFAM domains (rectangles), transmembrane helices (narrow vertically oriented rectangles) and signal peptides (at the N-terminus). Domain symbols and names are itemized in an online appendix (see Supplementary material). FSRs of orthologous families are written above gene names. Numbers below domains are rates of domains in PAM/BYr units.

We focus on the large multigene family of PTKs in greater detail and analyze the 14 domain architectures shown in Figure 4. Each of the domain architectures is present within 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 are membrane anchored and contain an extra-cellular ligand binding domain. 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 in Figure 4.

Family Specific Rates of orthologous families are written above gene names in Figure 4. 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 the domains out of the sequences and align them to the domain models using "hmmalign" [Edd98]. Finally, we apply the FSR estimator to the domain alignments. Domain rates are written below domain symbols in Figure 4. 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.

6 Summary and conclusion

We analyze evolutionary rates of protein families that comprise orthologs from man, fugu, fly, and worm. The assumption that the number of mutations per time unit is constant, the so called molecular clock hypothesis, allows to represent the evolution of a family by a rooted ultrametric phylogenetic tree where all leaves are equally distant to the root. In such a tree the edge lengths are proportional to the estimated number of mutation events and can be scaled with a rate parameter. The fact that a protein's evolutionary rate differs for different lineages, i.e., that the molecular clock does not hold in general, is accounted for by reconstructing an additive tree rather than an ultrametric one. We apply both tree models and a ML framework to estimate family specific evolutionary rates. A pregiven set of divergence times is used to relate measures of amino acid replacements to historical times.

We consider publicly available data of RNAi knockout experiments and high throughput 2-hybrid systems in *C. elegans* and establish relationships of family specific rates to the essentiality and the connectivity of proteins. We find that indispensable proteins are subject to strong purifying selection.

Finally, we analyze fast evolving extra-cellular proteins and the large multigene family of protein tyrosine kinases. For the latter we reveal that extra-cellular domains compared to their cytoplasmic counterparts are more divergent.

Supplementary material The whole set of orthologous families, together with alignments and ML trees, FSRs, tree length ratios, p_i-values of LRTs and 95%-bootstrap confidence intervals of FSRs is available available at http://speeds.molgen.mpg.de. Symbols and names of domains shown in Figure 4 are itemized in an online appendix available at http://speeds.molgen.mpg.de/gcb_appendix.

Acknowledgements Part of this work was supported by grants from the Bundesministerium für Forschung und Bildung, Germany, through its contribution to the Helmholtz Network for Bioinformatics. We thank Tobias Müller, Anja von Heydebreck, Sven Rahmann, Eike Staub, Antje Krause, and Thomas Manke for valuable discussions.

References

[CDKHK04] C. I. Castillo-Davis, F. A. Kondrashov, D. L. Hartl, and R. J. Kulathinal. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Research*, 14(5):802–811, May 2004.

[CPS+03] A. D. Cutter, B. A. Payseur, T. Salcedo, A. M. Estes, J. M. Good, E. Wood, T. Hartl, H. Maughan, J. Strempel, B. Wang, A. C. Bryan, and M. Dellos. Molecular correlates of genes exhibiting RNAi phenotypes in Caenorhabditis elegans. *Genome Research*, 13(12):2651–2657, Dec 2003.

[DP04] J. C. Davis and D. A. Petrov. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, 2(3), Mar 2004.

[Edd98] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA, 2004.

[HBB+02] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

[Hed02] S. B. Hedges. The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849, Nov 2002.

[HF01] A. E. Hirsh and H. B. Fraser. Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–1049, Jun 2001.

[HK05] M. W. Hahn and A. D. Kern. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-interaction Networks. *Molecular Biology and Evolution*, 22(4):803–805, 2005.

[KFD+03] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen, and J. Ahringer. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, 421(6920):231–237, Jan 2003.

[KFJ+04] E. V. Koonin, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, I. B. Rogozin, S. Smirnov, A. V. Sorokin, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, 5(2):R7, 2004.

[LAB+04] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657):540–543, Jan 2004.

[LCS+04] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, and P. Bork. SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, 32 Database issue:D142–4, Jan 2004.

[LSK+02] D. J. Lipman, A. Souvorov, E. V. Koonin, A. R. Panchenko, and T. A. Tatusova. The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, 2(1):20, Nov 2002.

[MSBP02] R. Mott, J. Schultz, P. Bork, and C. P. Ponting. Predicting protein cellular localization using a domain projection method. *Genome Research*, 12(8):1168–1174, Aug 2002.

[MSV02] T. Müller, R. Spang, and M. Vingron. A Comparison of Dayhoff's Estimator, the Resolvent Approach and a Maximum Likelihood Method. *Molecular Biology and Evolution*, 10(1):8–13, 2002.

[MV00] T. Müller and M. Vingron. Modeling Amino Acid Replacement. *Journal of Computational Biology*, 6:761–776, 2000.

[NEBvH97] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1–6, 1997.

[NR02] R. Nair and B. Rost. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, 18:S78–86, 2002.

[RSS01] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in–paralogs from pairwise species comparisons. *JMB*, 314(5):1041–1052, Dec 2001.

[SMD97] J. Stoye, V. Moulton, and A. W. Dress. DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *CABIOS*, 13(6):625–626, Dec 1997.

[SvHK98] E. L. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, pages 175–182, 1998.

[WF93] J. C. Wootton and S. Federhen. Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Computers and Biochemistry*, 17(3):149–163, 1993.

[WGP04] E. E. Winter, L. Goodstadt, and C. P. Ponting. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, 14(1):54–61, Jan 2004.

[WKH99] D. Y. Wang, S. Kumar, and S. B. Hedges. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 266:162–171, 1999.

[Yan96] Z. Yang. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *Journal of Molecular Evolution*, 42(5):587–596, May 1996.

[Yan97] Z. Yang. A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 13:431–439, 1997.