5 Phylogenetic trees from multiple genes

5.1 Inferring genome phylogenies

We turn to the problem of estimating a phylogenetic tree from an ensemble of protein families. Straight forward approaches to phylogeny construction from many genes scale up the traditional phylogenetic methods to apply to many families and derive a consensus. For example, evolutionary distances between pairs of orthologs can be averaged and used as input for a distance based reconstruction like Neighbor-Joining, a procedure that is sometimes referred to as *Individual Protein (IP) method* [Grishin *et al.*, 2000; Wolf *et al.*, 2001; Nei *et al.*, 2001]. Another useful method is the analysis of a large concatenated alignment [Hedges and Kumar, 2003]. This is sometimes done in conjunction with modeling positional evolutionary rates by a gamma distribution [Yang, 1996; Nei *et al.*, 2001]. However, the main limitation of this approach is that all genes are forced into one alignment under the assumption that the same evolutionary process acted on the genes. Several authors decide to restrict the analysis of concatenated alignments to a small set of well conserved families, e.g., to ribosomal proteins [Brown *et al.*, 2001; Wolf *et al.*, 2001].

Here, we propose two methods to reconstruct a phylogenetic tree from sets of families that incorporate the differences in evolutionary rates among families. The first method relies on the linear relationship observed in the scatter plots presented in Section 3.4. Fitting the slope of this relationship yields ratios of path lengths in the tree and thus allows to fit a tree fulfilling the ratios. We call this the *Path Length Ratio (PLR) method*. The second one makes use of estimated FSRs using them to rescale measured distances among proteins. Subsequent averaging yields an estimator for the divergence time of two species. This method can be seen as a variant of the IP method, we call it the *weighted Individual Protein (wIP) method*.

The PLR method implicitely assumes that the tree topology is known. The wIP method uses the tree topology for the estimation of Family Specific Rates. Yet, in cases where the tree topology is not known, family specific evolutionary rates can be estimated as tree lengths (see Section 3.6.1). The wIP method therefore is suited to estimate the edge lengths (or divergence times respectively) as well as the tree topology.

5.2 The Path Length Ratio (PLR) method

Given a topology for the phylogeny under study, the *Path Length Ratio (PLR)* method serves to estimate the edge lengths of the tree using a set of orthologous families. We exemplify it for the set of four organisms $\{H, F, D, C\}$ in our study. The method does not use a family specific evolutionary rate but instead relies on the scatter plots of orthologous distances.

We exploit the scatter plots that were presented in Section 3.4 for estimating ratios ρ_i among edge lengths $\tau_1, ..., \tau_5$ of the organismal phylogeny (see Figure 5.1). To this end, a regression line through the origin is fitted to the data points of each scatter plot. We perform total least squares regressions by taking the variances of the distance estimates into account as described in Section 3.4. In this way, we derive 15 values ρ_i, $i = 1, ..., 15$ for the slopes of the regression lines in the scatter plots.

To name a concrete example, suppose that the *HF-DC* scatter plot defines a path length ratio ρ_{HF-DC}. Using the notations of the tree in Figure 5.1, the distance of H and F corresponds to $\tau_1 + \tau_2$ and the distance of D and C corresponds to $\tau_4 + \tau_5$. Equating these terms we obtain $\rho_i = \rho_{HF-DC} = (\tau_4 + \tau_5)/(\tau_1 + \tau_2)$. Converting this to a homogeneous system of 15 linear equations will in general allow only for the trivial solution of all edge lengths being 0. Instead we define error terms of the form

$$e_i = (\tau_4 + \tau_5)/(\tau_1 + \tau_2) - \rho_i.$$

Note that for each scatter plot the term needs to be written out seperately. This will yield 15 equations. Solving the equations will define a one dimensional manifold of trees all of them approximating the given path length ratios. By fixing the historical time of an edge one tree is selected. We proceed to minimize the quadratic error functional

$$E = \sum_{i=1}^{15} e_i^2 \to \min$$

The minimization of this quadratic functional is achieved numerically [Lagarias *et al.*, 1998].

5.3 The weighted Individual Protein (wIP) method

The Individual Protein method averages over the distances of orthologous pairs. If protein families each obey their own family-specific evolutionary rates, this procedure appears to be non-optimal. Instead, before averaging, the measured distances between pairs should be re-scaled by the applicable FSR. In more formal terms, for each pair

of organisms (k, l) and each family i we compute

$$\Theta^i_{kl} = \hat{t}^i_{kl}/\hat{\lambda}_i$$

with $\hat{t}^i_{kl}$ denoting the evolutionary distance in orthologous family i being included in the set of all considered orthologous families $\mathcal{G}$, $i \in \mathcal{G}$. The average

$$\langle\Theta\rangle_{kl} = \frac{1}{|\mathcal{G}|}\sum_i \Theta^i_{kl}$$

can then be used as a new estimate for the divergence time of the organisms and can be input to a traditional distance based tree-building algorithm like, e.g., Neighbor-Joining.

5.4 Results

For the species under study here the topology of the tree is not under discussion. However, the methods introduced above may serve to re-estimate the divergence times on our reference tree using a large set of orthologous families. Clearly, in order to obtain reliable results, orthologous families should be used that do not display rate heterogeneity. We use the p_i-values of the likelihood ratio test (see Section 3.5.3) to rank the protein families according to how well they fit the assumption of rate-constancy. To do this, we define the set $\mathcal{S}_p$ to consist of those families for which $p_i > p$. Clearly $\mathcal{S}_0$ contains all orthologous families (3640), whereas in $\mathcal{S}_{0.05}$ this number is reduced to 888. We vary the values of p from 0.95 to 0 and obtained the corresponding sets of families $\mathcal{S}_p$. Recall that these sets are nested and include more and more orthologous families which do not perfectly obey rate constancy.

We apply both the PLR and the wIP method to compute edge lengths for each of the sets $\mathcal{S}_p$. We use Neighbor-Joining implemented in PHYLIP [Felsenstein, 1993] to obtain edge lengths in the wIP method. The height of the bars in Figure 5.2 give the

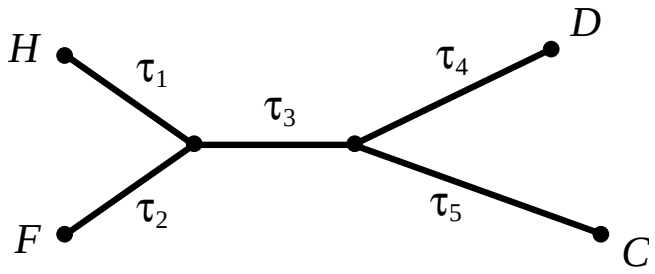

Figure 5.1: Tree topology for the species *Homo sapiens (H)*, *Fugu rubripes (F)*, *Drosophila melanogaster (D)* and *Caenorhabditis elegans (C)*.

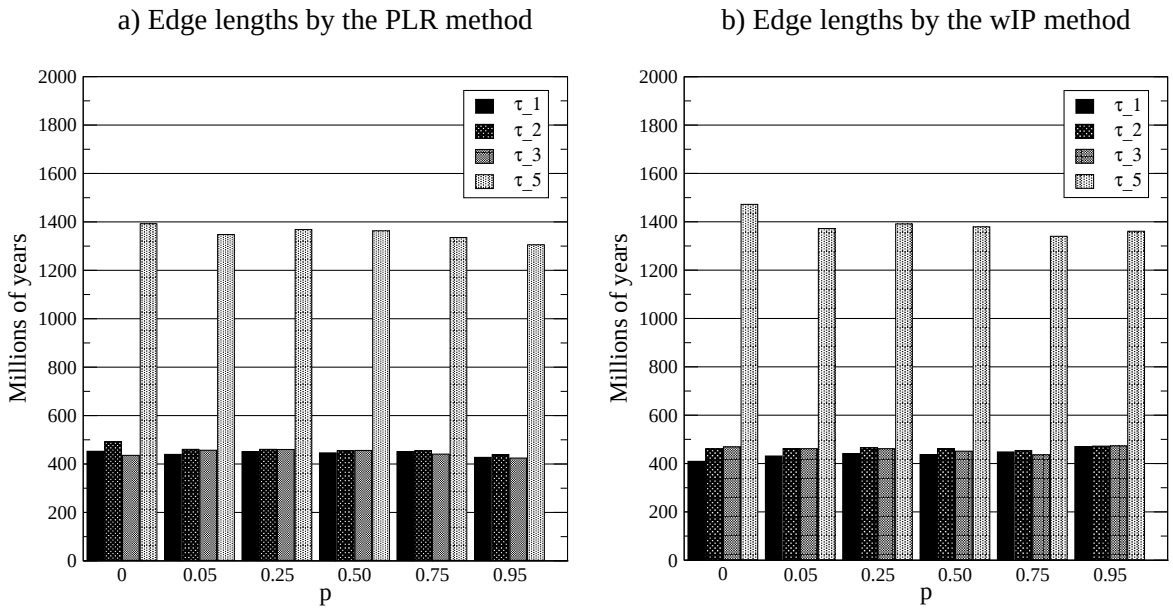

Figure 5.2: Edge lengths τ_1, τ_2, τ_3 and τ_5 (see Figure 5.1) of trees reconstructed by the wIP and the PLR method and on different sets $\mathcal{S}_p$-values. Trees were calibrated with respect to the divergence time τ_4 of *Drososphila melanogaster* and *Homo sapiens* that is given in the reference tree (see Figure 5.3a).

edge lengths τ_1, τ_2, τ_3 and τ_5 for different values of p when fixing the edge length τ_4 to the one of the reference tree, that is to the given divergence time of *Drososphila melanogaster* and *Homo sapiens (H)* (see Figure 5.3a). As long as a set of rate constant familes is selected, the resulting trees vary very little. When dropping the restriction and going to the complete set $\mathcal{S}_0$ of orthologous families, the edge to *C. elegans* is stretched and the difference in the edges τ_1 and τ_2 to *Homo sapiens* and *Fugu rupripes* becomes larger. This behaviour is indepenent of whether PLR or wIP is used.

Figure 5.3b shows a tree that reflects the edge lengths of the trees we reconstructed for sets $\mathcal{S}_p$ and $0.05 < p < 1$. The most striking difference to the reference tree in Figure 5.3a is the shorter inner edge. After calibrating the edge lengths to million years with respect to the divergence time of fly and man given in the reference tree, the inner edge in the time between the divergence of fly and the split between man and fish is 452 million years. This is considerably shorter than the 543 million years in the reference tree. As a consequence the newly constructed tree is not ultrametric implying that the molecular clock assumption is not valid for the four species. Indeed, a recent study confirms that evolutionary rates in vertebrates were significantly decreased with respect to invertebrates [Peterson *et al.*, 2004].

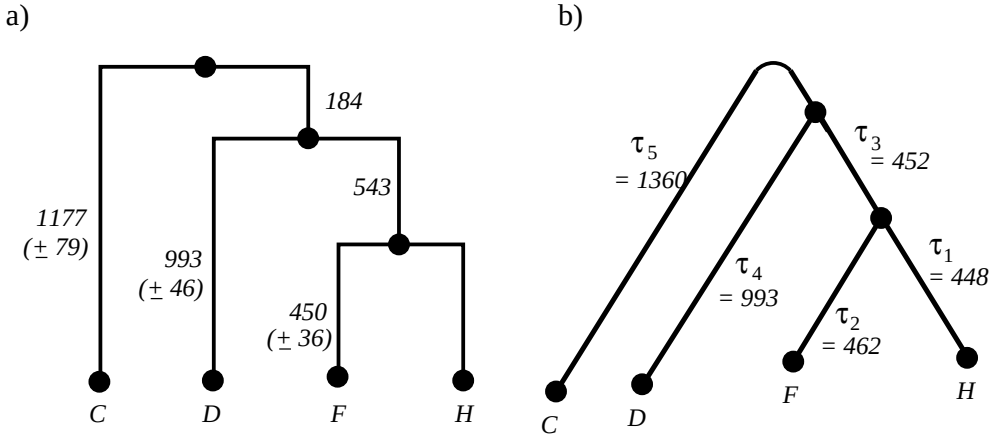

Figure 5.3: Trees representing organismal phylogenies. a) Ultrametric tree taken from [Wang *et al.*, 1999; Hedges, 2002]. Numbers indicate divergence times, the units are millions of years. b) Tree and edge lengths reconstructed by the PLR and the wIP method. Numbers were obtained by calibration with respect to the divergence time τ_4 of *Drososphila melanogaster (D)* and *Homo sapiens (H)* in the reference tree shown in a). The tree is unrooted and non-ultrametric. The molecular clock assumption is not valid and the rates of evolution are supposed to have changed in certain lineages. Since a root position cannot be inferred, we decide to represent the edge to *C. elegans* as an arcuated line.