World Scientific
www.worldscientific.com

Deviations from Ultrametricity
in Phage Protein Distances

Chad Wagner[1], Anna Salamon[1,2], Robert A. Edwards[3,4],
Forest Rohwer[5], and Peter Salamon[1]

[1]*Department of Mathematics and Statistics, San Diego State University,
San Diego, CA, 92182, USA*

[2]*Department of Philosophy, University of California, La Jolla, CA 92093, USA*

[3]*Department of Computer Science, San Diego State University,
5500 Campanile Dr, San Diego, CA 92182, USA*

[4]*Mathematics and Computer Science Division, Argonne National Laboratory,
9700 S. Cass Ave, Argonne, IL 60439, USA*

[5]*Department of Biology, San Diego State University,
5500 Campanile Dr, San Diego, CA 92182, USA*

Abstract. Distances in biological databases are known not to be ultrametric. *Deviations* from ultrametricity can however reveal useful features of biodata. In the present study we examine deviations from ultrametricity of the distances between known phage proteins quantified in two senses: (1) the failure of triangles to be isosceles and (2) failure of every point to be the center of any sphere in which it resides. The deviations from these two ultrametric properties undergo qualitative changes as a function of the distance. Below we describe these changes and how they can be observed. We further argue that the distances at which the qualitative changes take place reveal intrinsic scales in the dataset. Such scales are important for choosing threshold values of the distance in various algorithms and reveal natural chunking of the data that can be used to decide clade levels in phage phylogeny.

1. Background

Phages are the most abundant biological entities on earth [1], and they are major players in biogeochemical cycles. Relatively little is known about them because most phage are recalcitrant to culturing using standard approaches. Elucidation of their phylogeny has proved difficult because they do not have any conserved proteins that can be found in all phage. Distance-based approaches to understanding phage phylogeny have been moderately successful [2, 3]. Such approaches can benefit from the information that protein distances show intervals of qualitatively different behaviour as a function of distance. For example, the Phage Proteomic Tree [2] counted all protein

similarities up to a threshold of 10 units[a]; the present paper shows that a smaller cutoff may have been more judicious.

An ultrametric space is a set of points with a distance function (a metric) that satisfies the strong triangle inequality [4, 5]. Specifically, letting $d(x, y)$ denote the distance between points x and y, an ultrametric space is one in which

$$d(x, z) \leq \max\{d(x, y), d(y, z)\} \tag{1}$$

for all points x, y, and z in the space. Ultrametric spaces have a number of properties that seem somewhat surprising to geometric intuition gleaned from Euclidean geometry. They are however, the natural geometry for indexed hierarchies. In fact, there is a one-to-one correspondence between ultrametric spaces and indexed hierarchies [5]. Any ultrametric space can be displayed as the lowest layer in a tree with the vertical distance representing clustering at various levels (see Fig. 1) [5]. The time since a common ancestor defines an ultrametric distance on the set of current species. The use of ultrametric spaces in taxonomy was popular circa 1960–80 [6, 7], but eventually declined [8, 9, 10] due primarily to the realization that distances calculated from DNA similarity are not ultrametric. Here we show that despite this fact, their nearly ultrametric structure can reveal intrinsic scales.

The two surprising properties of ultrametric spaces that are of relevance to our calculations below are:

Property 1: All triangles in an ultrametric space are isosceles, with the two longer sides having equal lengths.

Property 2: Two spheres that have the same radius and have any points in common, have all their points in common.

These properties are illustrated in Fig. 1. We will examine deviations from these two properties in the sections below.

2. Results and Discussion

2.1. THE DISTANCE DISTRIBUTION

Protein distance data were calculated for all 19,619 proteins in the 410 complete phage genomes in GenBank as of November 2004. Of the $\binom{19619}{2} \approx 2{\cdot}10^8$ pairs of proteins, only about $6{\cdot}10^5$ pairs showed significant similarity; the rest were assigned a distance of infinity. A large number of infinite distances are expected given the diverse functions these proteins perform, and that they are not related to one another. The histogram of all finite distances found in this manner is shown in Fig. 2. Note that the data consists of two peaks, a sharp one near zero distance and a broad one centered around $d = 1.8$.

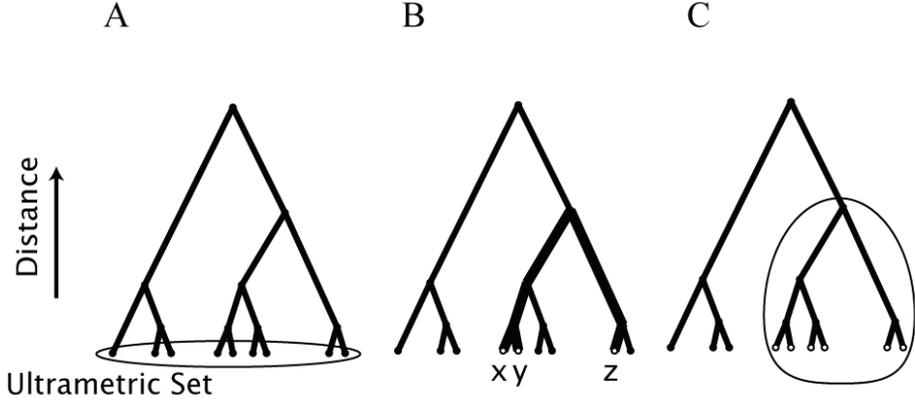[a]See more careful definition of the distance calculation below.

Fig. 1: The lowest layer of nodes in a tree with distance defined by height required to reach from one vertex to another define an ultrametric set (panel A). In panel B, the edges used to calculate the distances between three points (x, y, and z) in the ultrametric set are darkened showing that the lengths of the two longest sides ($d(x,z)$ and $d(y,z)$) are equal. In panel C, a sphere containing six points is shown. This same sphere is obtained as the sphere with radius defined by the highest common ancestor and centered at any of the six points.

2.2. Deviations from isosceles among triangles

For the purpose of this calculation, a triangle is defined as three proteins with finite distances between each pair. Of the potentially $\binom{19619}{3} \approx 10^{12}$ triples, only about $2 \cdot 10^6$ of them were triangles in this sense. Each triangle was assigned a deviation Δ, which was defined as the absolute value of the difference between the lengths of the two longest sides. The Δ values were collected into bins based on the length of the longest side in the triangle and using bins of size 0.01. The mean and standard deviation of the Δ values in each bin are shown in Fig. 3. Note four distinct regions of qualitatively different behaviour are approximately as follows:

$$
\begin{aligned}
0 &\leq d \leq 0.2 \\
0.2 &< d \leq 0.95 \\
0.95 &< d \leq 2.0 \\
2.0 &< d < +\infty
\end{aligned}
$$

where d is the average of the two longest sides. The larger scatter in some regions of the figure is NOT due to sample size; all bins contain the same
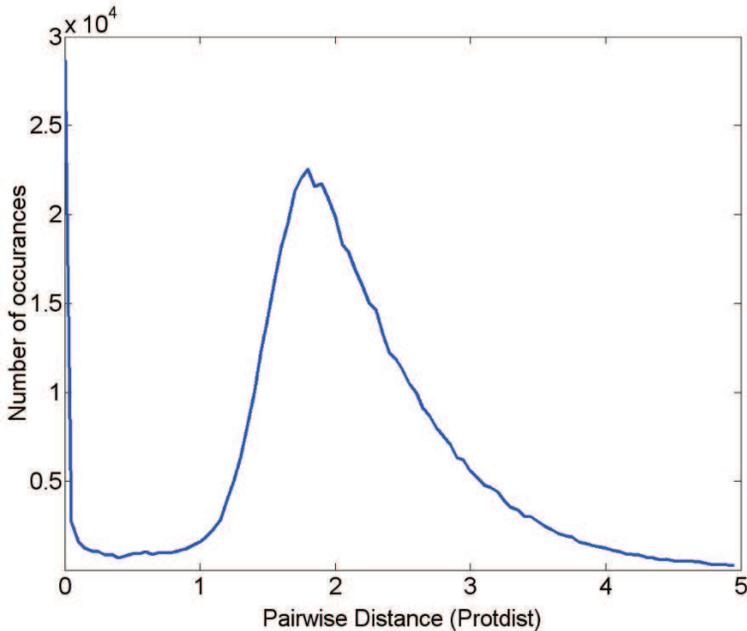
Fig. 2: Histogram of all finite protein-protein distances from the database of all phage protein distances.

number of points and the pattern is seen with other random sets of this many points for any of the bins. While the thresholds are approximate, the differences between the regions are evident. Note that around $d = 0.95$ the scatter increases significantly. This will have implications for our model landscape described below.

2.3.   DEVIATIONS FROM SPHERE OVERLAPS

The second property of ultrametric data is that spheres centered at points closer than their radius coincide. The algorithm developed to test this property is summarized in Fig. 4. Briefly a radius r was fixed, and for each protein $i, i = 1, \ldots, N$, the characteristic function

$$P_i(k) = \left\{ \begin{array}{ll} 1 & d(p_i, P_k) < r \, , \\ 0 & \text{otherwise} \end{array} \right. \tag{2}$$

of the sphere of radius r centered at that protein was calculated. The correlations between the functions P_i and P_j were calculated, and collected into bins based on the value of $d(i, j)$. The mean and standard deviation of the correlations were plotted as a function of distance. The resulting plots for various values of the radius r are shown in Fig. 5. For a perfect ultrametric structure, these plots would stay equal to one for $d < r$ and jump to a value
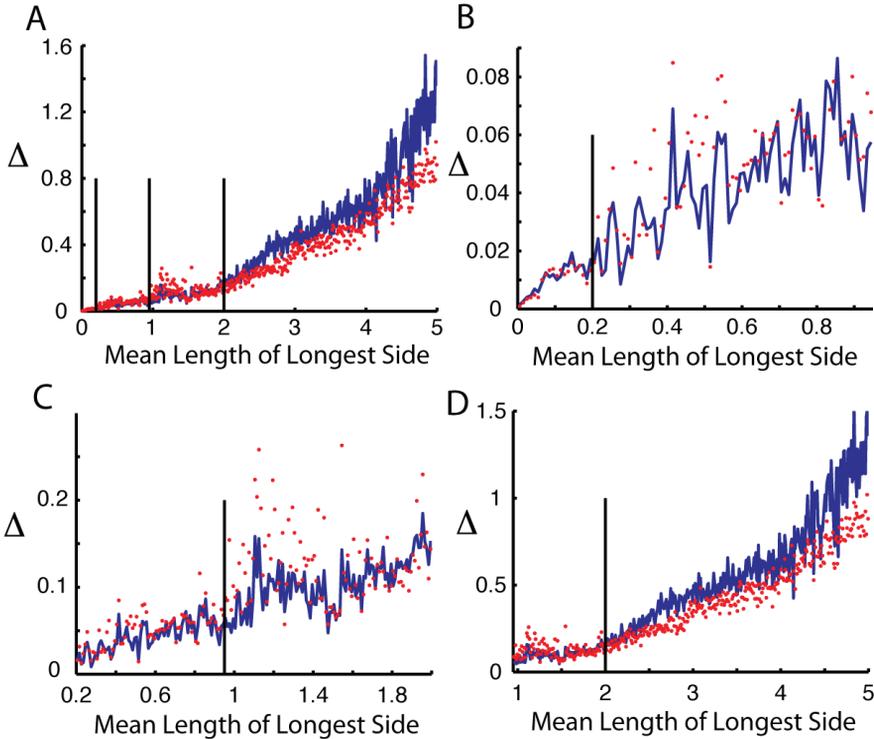
Fig. 3: Mean (blue line) and standard deviation (red dots) of the deviations Δ from isosceles, measured as the difference between the two longest sides. Panel A shows the data over the full range with vertical lines indicating the positions of qualitative changes. The changes are further illustrated in panels B-D, each of which shows a close-up of two adjacent regions.

near zero for $d > r$. Indeed, for small r, this is approximately the behaviour seen. Once r passes about 1.0, this behaviour erodes into a monotonic decrease with a corner around $d = 0.95$ and with smaller and smaller jumps at $d = r$. By $r = 2.2$, the jump is hardly visible. We again note the pivotal role of $d = 0.95$.

2.4.   PHAGE PROTEIN DISTANCE LANDSCAPE

The previous two sections examined deviations from ultrametricity for triangles and spheres in the set of phage proteins. Both showed approximately ultrametric behaviour for small distances. Both showed qualitative changes in how they deviated from the ultrametric structure when the distance passed the values $d = 0.95$ and $d = 2.0$. Deviations in triangles also showed a change around $d = 0.2$. In retrospect, all three of these threshold values of d are

Mean Sphere Correlation Algorithm
1. Fix a radius r.
2. Initialize to zero a matrix P of the "distance less than
   r relation" and reset the value of $P(i,j)=1$ if $d(i,j)< r$.
3. Initialize 100 lists indexed by the distance interval 0
   to 5 broken into bins of size 0.05.
4. Loop over proteins i,j, $i<j$.
   i. Find the correlation between the i-th and j-th
       columns of P.
   ii. add this correlation, corr(P_i, P_j), to the bin
       corresponding to $d(i,j)$.
5. For each of the bins calculate the mean and standard
   deviation of the correlations in that bin.
6. Plot the resulting mean correlation, and mean ± one
   standard deviation versus distance.
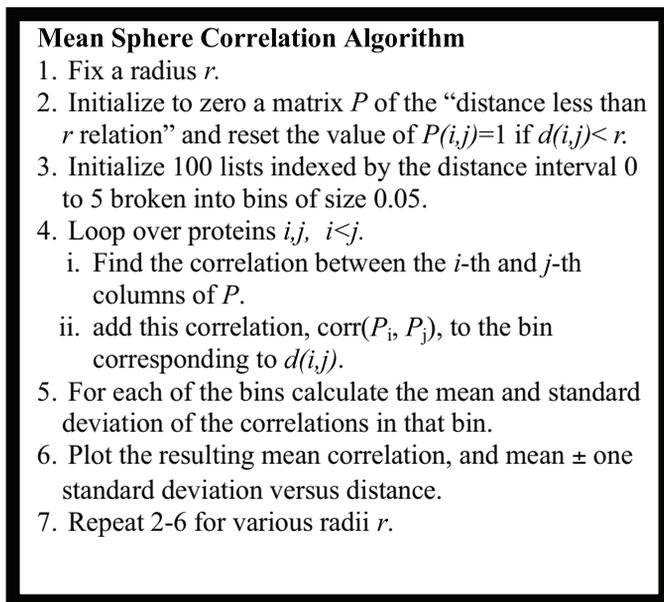7. Repeat 2-6 for various radii r.

Fig. 4: Algorithm for calculating sphere-sphere correlations at various radii.
Note that one iteration of steps 2-6 results in one panel of Fig. 5.

visible also in the original histogram (Fig. 2). Although the protein-protein
distances do not even define a proper distance function[b], these distances are
similar enough to an ultrametric space that measuring deviations has re-
vealed useful thresholds separating distinct regions of qualitative behaviour.

As a *tentative* interpretation, we advance an approximate local picture
of the phage protein distance landscape shown in Fig. 6. We posit that this
landscape consists of small, nearly ultrametric clusters with radius less than
0.5. These clusters are relatively far apart, being linked at distances greater
than 1 with significantly more deviations from ultrametricity.

This tentative geometrical structure leads to testable hypotheses. One
characteristic of all the local neighbourhoods is that when we move from
neighbours at a distance less than 1 to a distance greater than 1, the next
distance is relatively far away. As partial confirmation of this approximate
geometrical structure, we plot in Fig. 7 the mean distance to neighbours as
a function of normalized rank. The calculation behind the plot starts by
ranking all neighbours of a protein according to distance from the protein.
These ranks are then rescaled by the number of neighbours the protein has
at a distance less than $d = 0.95$. The average distance at each relative

---

[b]One that is nonnegative, symmetric, and satisfies the (weak) triangle inequality. The
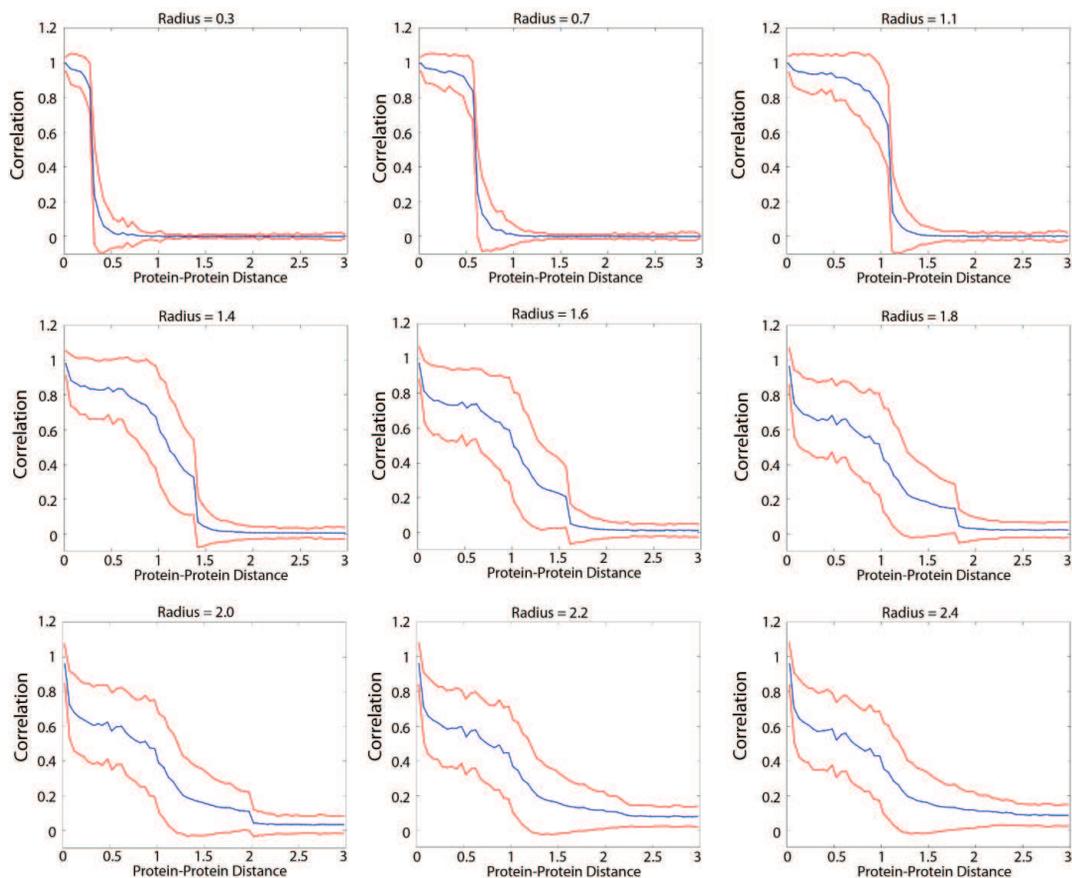data contains several violations of the (weak) triangle inequality.

Fig. 5: The three curves show the mean (blue) and the mean ± one standard deviation (red) of the correlation between characteristic functions of spheres plotted as a function of the distance between sphere centers for various radii.

rank is then plotted. The idea behind the calculation is that when we cross the threshold at $d = 0.95$, the distance jumps rather significantly, an idea confirmed by the observed jump in mean distance from around 0.5 to around 1.3. Note that this is only the mean structure of the phage protein landscape and is at best approximate.

There are several likely causes for such a landscape. For example, differential evolutionary pressures and rates of evolution, horizontal gene transfer, and the less than perfect correspondence between measured and ideal distances may all contribute to the observed deviations from ultrametricity.
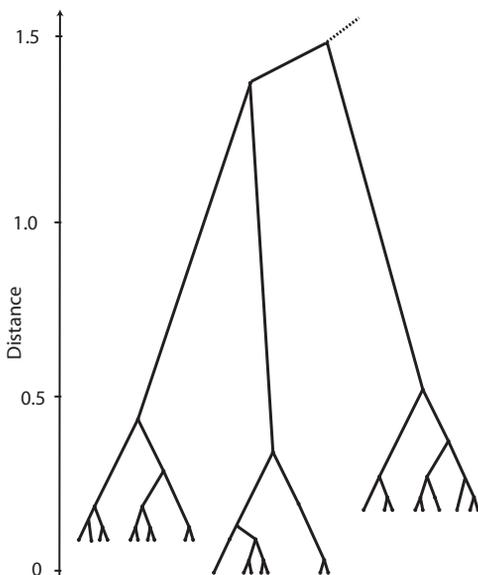
C. Wagner *et al.*

Fig. 6: Tentative local structure of the phage protein distance landscape.

3.   Methods

The "proteins" are ORFs identified computationally from the 410 completely
sequenced phages in GenBank as of November 2004. All pairs of proteins
were compared for similarity using the heuristic BLAST [11] algorithm with
permissive parameters. The similar pairs of proteins were aligned in parallel
using CLUSTALW [12] on a 36-node LINUX cluster and the distance between
them was calculated using the ProtDist program available in J. Felsenstein's
PHYLIP package [13, 14] which implements the algorithm of Jones-Taylor-
Thornton [15]. The resulting matrix of pairwise distances was used for all
our calculations. Calculations using this matrix were performed with Matlab
version R2007A.

4.   Conclusions

Protdist values among phage proteins show four different regimes of be-
haviour. The values of the distance where qualitative changes occur can
be useful for the selection of threshold values for which pairs to count in
any algorithm based on protein-protein distances. These threshold values
are also useful in determining the natural levels of chunking, and associated
clade definitions for phage phylogeny [2].

   Although phage protein distances do not satisfy the formal definition
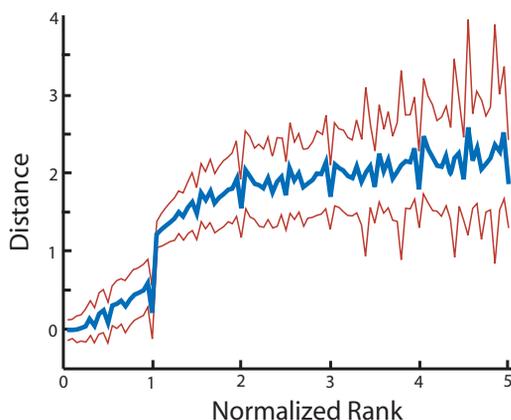of an ultrametric space, *approximate* ultrametric structure on biodata can

Fig. 7: The blue curve shows mean distance to neighbours as a function of normalized rank. Neighbours of each protein were ranked by distance, and these ranks were normalized by dividing by the number of neighbours at a distance less than 0.95. The red curves are displaced $+/-$ one standard deviation from the mean.

be useful for gauging biologically relevant differences. The scales revealed by such studies can be a guide to appropriate choices of natural scales for chunking in biological databases.

Acknowledgements

Bibliography

[1] R. A. Edwards, F. Rohwer, *Viral metagenomics.*, Nat. Rev. Microbiol. **3**, 504 (2005).

[2] F. Rohwer, R. Edwards, *The Phage Proteomic Tree: a genome-based taxonomy for phage*, J. Bacteriol. **184**, 4529 (2002).

[3] M. L. Pedulla, M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. J. Jacobs, R. W. Hendrix, G. F. Hatfull, *Origins of highly mosaic mycobacteriophage genomes*, Cell **113**, 171 (2003).

[4] See the Wikipedia entry http://en.wikipedia.org/wiki/Ultrametric_space.

[5] R. Rammal, G. Toulouse, M. Virasoro, *Ultrametricity for physicists*, Rev. Mod. Phys. **58**, 765 (1986).

[6] J. Benzécri, *et al.*, *L'analyse des donnees. 1.La taxinomie*, Dunod 1973.

[7] J. Hartigan, *Representation of Similarity Matrices by Trees*, Journal of the American Statistical Association **62**, 1140 (1967).

[8] P. Landry, F. Lapointe, J. Kirsch, *Estimating Phylogenies from Lacunose Distance Matrices: Additive is Superior to Ultrametric Estimation*, Molecular Biology and Evolution **13**, 818 (1996).

[9] D. Penny, L. Foulds, M. Hendy, *Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences*, Nature **297**, 197 (1982).

[10] A. Wilson, *The Molecular Basis of Evolution*, Scientific American **253**, 164 (1985).

[11] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Research **25**, 3390 (1997).

[12] J. Thompson, D. Higgins, T. Gibson, *et al.*, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*, Nucleic Acids Res **22**, 4673 (1994).

[13] J. Felsenstein, *PHYLIP-Phylogeny Inference Package (Version 3.2)*, Cladistics **5**, 164 (1989).

[14] J. Felsenstein, *PHYLIP (Phylogeny Inference Package), version 3.57 c*, Seattle, University of Washington, 1995.

[15] D. Jones, W. Taylor, J. Thornton, *The rapid generation of mutation data matrices from protein sequences*, Bioinformatics **8**, 275 (2003).