A COVARIANCE MATRIX INVERSION PROBLEM ARISING FROM THE CONSTRUCTION OF PHYLOGENETIC TREES

TOM M. W. NYE, BRAD J. C. BAXTER AND WALTER R. GILKS

Abstract. We describe an efficient algorithm for the inversion of covariance matrices that arise in the context of phylogenetic tree construction. Phylogenetic trees describe the evolutionary relationships between species, and their construction is computationally demanding. Many approaches involve the symmetric matrix of evolutionary distances between species. Regarding these distances as random variables, the corresponding set of variances and covariances form a rank-4 tensor, and the inner-product defined by its inverse can be used to assign statistical scores to candidate trees. We describe a natural set of assumptions for the phylogenetic tree under construction, and show how under these assumptions the covariance tensor for a tree with n leaves can be inverted in $O(n^2)$ operations. In addition to presenting the inversion algorithm, we hope this article will open algebraic and computational problems from the field of phylogeny to a wider audience.

1. Introduction

Suppose we are given a set of n species and an $n \times n$ symmetric matrix of random variables (d_{ij}) representing the evolutionary distances between them. In this paper we show how simple assumptions on the tree of evolutionary relationships between species gives rise to a covariance matrix which essentially has the form

$$\mathrm{Cov}(d_{ij}, d_{kl}) = \frac{1}{2}B(\delta_{ik}\delta_{jl} + \delta_{jk}\delta_{il}) + C^2 + \frac{1}{2}Ca_j(\delta_{jk} + \delta_{jl}) + \frac{1}{2}Ca_i(\delta_{ik} + \delta_{il}), \quad (1)$$

for $i \neq j$ and $k \neq l$, where B and C are constants, $a \in \mathbb{R}^n$, and δ_{ij} is the Kronecker delta. The great advantage of our model is that covariance matrices of this form can be inverted in $O(n^2)$ operations, via the Sherman–Morrison–Woodbury formula (see, for instance, p. 51 of [6]).

In order to motivate this inversion problem it is necessary to present some background from the field of phylogeny. While this area might be unfamiliar to the reader, we hope that the mathematical problems it raises will be of interest. The rest of this section gives a very brief introduction to phylogeny, before we describe the origin of equation (1) in Section 2. We show how the covariance tensor can be inverted algebraically in Section 3, and present the inversion algorithm in Section 4. Readers principally concerned by the linear algebraic details of the inversion algorithm rather than the bioinformatic background may skip directly to Section 3.

Evolutionary relationships between species can be represented by a tree: the leaf nodes represent extant species, interior nodes represent ancestral species, and the branch lengths indicate the extent to which species have diverged. Such trees are referred to as *phylogenies*. One kind of tree is illustrated in Figure 1. Here the branch lengths specify the time since divergence of species, and it is this type of tree we will consider throughout this paper.

There are a range of different statistical methods available for inferring the phylogeny of a set of species given their DNA sequences, or subsequences, which might

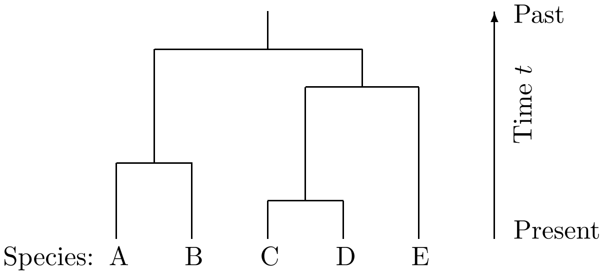
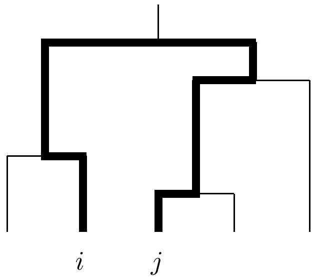


FIGURE 1. A typical phylogenetic tree



$$\mathbb{E}\left[d_{ij}|T\right] \propto |t_{ij}| = \text{length of highlighted path}$$

FIGURE 2. Expected evolutionary distance between species is proportional to path length in T

be single genes, subsets of genes, or even entire genomes. All of these methods are based on the fundamental idea that species with similar DNA sequences are more closely related than species for which the sequences have diverged, as mutations in sequence accumulate with time [4]. One class of methods, the so-called *distance-based* approaches, constructs a matrix of evolutionary distances, or distance functionals, between species that hopefully summarizes the information contained in the full set of sequences [3, 7, 5, 2]. The evolutionary distance between two species typically measures the number of letter changes in the DNA sequence, and many different distance functionals exist. Distance-based methods take the matrix of distances between extant species and hence infer the topology and branch lengths of the underlying phylogenetic tree. They have the advantage of being relatively fast, and therefore suitable for large problems, but are less suitable when the set of genes under investigation has diverged widely.

2. Origin of the covariance tensor

The inversion problem studied in this paper arose from a novel distance-based phylogenetic method developed by two of the authors (W. Gilks and T. Nye). Given a set of species and the matrix of evolutionary distances between them, our approach resconstructs the underlying phylogenetic tree, denoted T, using the following set of assumptions. Given two points $i, j \in T$, let t_{ij} denote the path in T between the points and let $|t_{ij}|$ denote the length of this path (or strictly speaking, its vertical component) as drawn in Figure 2. We assume that the evolutionary distance between two genes is, *on average*, directly proportional to the time since divergence. Denoting the distance between nodes i and j by d_{ij} we therefore obtain

$$\mathbb{E}\left[d_{ij} \mid T\right] = \mu|t_{ij}| \tag{2}$$

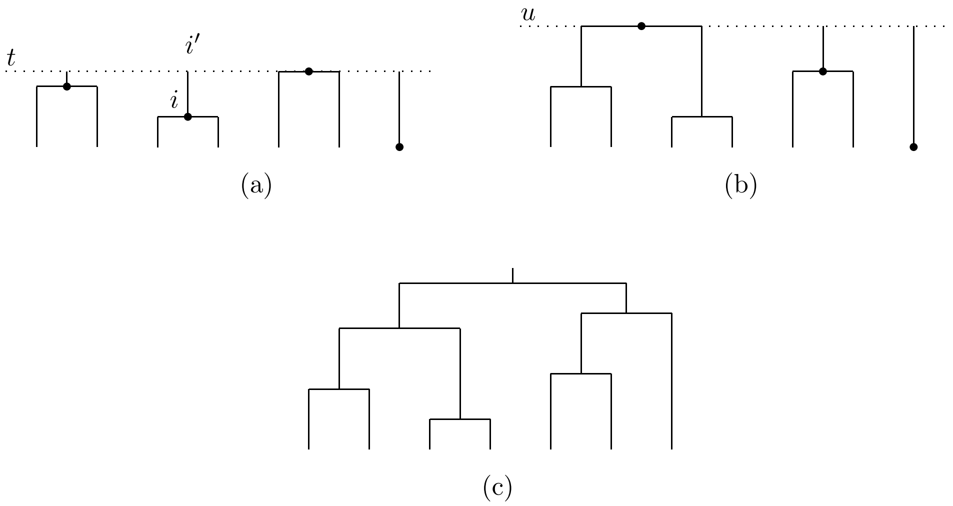


FIGURE 3. Tree construction. (a) A generic point in construction: the tree is complete up to time t. The set of hanging nodes H_t is highlighted with circles. (b) The next step in the construction obtained from (a). Construction is complete up to a time $u > t$. The set of hanging nodes H_u is highlighted. (c) The finished tree.

for some constant μ. As usual, the notation $\mathbb{E}[X \mid A]$ denotes the expectation of the random variable X conditional the occurrence of event A.

Our approach makes an additional assumption on the variances and covariances of the distances d_{ij}:

$$\operatorname{Cov}\left[d_{ij}, d_{kl} \;\middle|\; T\right] = \nu |t_{ij} \cap t_{kl}| \tag{3}$$

for some constant ν. In other words, the covariance of two distances d_{ij} and d_{kl} is proportional to the length of the shared path between these genes on the underlying phylogenetic tree. Direct calculation using equations (2) and (3) provides

$$\mathbb{E}\left(d_{ik} - d_{ij} - d_{jk}\right) = 0$$

and

$$\mathbb{E}\left[\left(d_{ik} - d_{ij} - d_{jk}\right)^2\right] = 0,$$

which imply the relation

$$d_{ik} = d_{ij} + d_{jk}, \tag{4}$$

where j can be any node on the path between nodes i, k in T. Thus the observed distances between extant genes arise from a distorted version of the underlying phylogenetic tree. One way to deform the tree T in this way would be via a gamma process on each branch; however, such probabilistic details are not our main concern here.

Equation (3) defines the covariance structure when the underlying phylogenetic tree T is known. However, our estimate of T is built up as a sequence of partially constructed trees, as illustrated by Figure 3. A generic stage of the construction is shown in Figure 3(a). This consists of an estimate of T constructed back as far as some time t, which we denote T_t. The covariance tensor defined in Equation (1) arises from considering the set of nodes descended directly from time t with no bifurcation. We refer to these as 'hanging nodes', as suggested by the appearance in Figure 3, and the set of such nodes is denoted H_t. These nodes are highlighted by black circles in Figure 3(a). Given a node $i \in H_t$ we use the notation i' to denote the ancestor of i at time t, and let t_i denote the time of node i. The length of the line segment between i and i' in Figure 3(a) is therefore $t - t_i$.

Given $i, j, k, l \in H_t$ such that $i \neq j$ and $k \neq l$, the additivity condition (4) gives

$$\operatorname{Cov}\left[d_{ij}, d_{kl} \;\middle|\; T_t\right] = \operatorname{Cov}\left[d_{ii'} + d_{i'j'} + d_{j'j}, d_{kk'} + d_{k'l'} + d_{l'l} \;\middle|\; T_t\right]. \tag{5}$$

The distances $d_{ii'}$ and $d_{kk'}$ have a covariance of zero (when $i \neq k$) because the corresponding tree branches have no overlap (ie. the right-hand side of Equation (3) is zero). This applies to the other indices as well, so expanding the right-hand side of Equation (5) gives

$$\begin{aligned}\operatorname{Cov}\left[d_{ij}, d_{kl} \;\middle|\; T_t\right] = (\delta_{ik} + \delta_{il})\operatorname{Var}[d_{ii'} \mid T_t] \\ + (\delta_{jk} + \delta_{jl})\operatorname{Var}[d_{jj'} \mid T_t] + \operatorname{Cov}[d_{i'j'}, d_{k'l'} \mid T_t].\end{aligned} \tag{6}$$

The first two terms of this expression can be obtained using Equation (3):

$$\operatorname{Var}[d_{ii'} \mid T_t] = \nu(t - t_i), \quad \text{and} \quad \operatorname{Var}[d_{jj'} \mid T_t] = \nu(t - t_j).$$

The final term of Equation (6) depends on the paths joining i' to j' and k' to l'. However, these paths lie in the part of the tree that has not yet been estimated (ie. above the dotted line in Figure 3(a)). By symmetry, since we condition only on T_t, the final term of Equation (6) depends only on whether the two paths $t_{i'j'}$ and $t_{k'l'}$ share common terminal nodes: it adopts three different values according to whether the paths share zero, one, or two terminal nodes. This gives

$$\begin{aligned}\operatorname{Cov}\left[d_{ij}, d_{kl} \;\middle|\; T_t\right] = \nu(t - t_i)(\delta_{ik} + \delta_{il}) + \nu(t - t_j)(\delta_{jk} + \delta_{jl}) \\ + c_0 + c_1(\delta_{ik} + \delta_{il} + \delta_{jk} + \delta_{jl}) + c_2(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\end{aligned} \tag{7}$$

for some constants c_0, c_1, c_2. This equation is now in exactly the same form as Equation (1), and it is this tensor that is used to score different partially constructed trees.

Construction of the estimated phylogeny is carried out in the following way. Given a partially constructed tree T_t, at each stage we propose a tree T_u $(u > t)$ by joining together two nodes from H_t with a new node at time u. One such tree is shown in Figure 3(b). This tree T_u is assigned a χ^2 statistic using the inverse of the covariance tensor (7), and the time u is chosen to optimize this score. Every possible tree T_u derived from T_t by joining two nodes together is scored in this way, and any tree with minimal u is taken as the new estimate. The process starts from the set of leaf nodes and ends when the entire tree has been estimated. Of course, in order to evaluate the score for each tree it is necessary to invert the covariance tensor many times, leading to our inversion problem.

3. Inverting the covariance tensor

We start by defining some notation. Let $V = \mathbb{R}^n$ be equipped with the standard basis $\{e_i : i = 1, \ldots, n\}$ and define the vector of all ones

$$e = \sum_{i=1}^{n} e_i,$$

If W denotes the vector space of symmetric $n \times n$ real matrices, then the covariance matrix defined by equation (1) is given by

$$H = B \times \mathrm{id} + C^2\theta \otimes \theta + C\alpha \otimes \theta + C\theta \otimes \alpha$$

where $\theta = e \otimes e$ and $\alpha = \mathrm{diag}(a_1, \dots, a_n)$. Incidentally, we shall move between the equivalent notations $u \otimes v$ and uv^T, for $u, v \in \mathbb{R}^n$, as appropriate. If we view $\mathbb{R}^{n\times n}$ as the tensor product space $\mathbb{R}^n \otimes \mathbb{R}^n$, then the inner product

$$\langle a \otimes b, c \otimes d \rangle = \langle a, c\rangle \, \langle b, d\rangle, \qquad a, b, c, d \in \mathbb{R}^n,$$

implies that

$$\langle M_1, M_2 \rangle = \sum_{j,k=1}^{n} M_1(j,k) M_2(j,k).$$

In other words, our inner product is simply the Frobenius inner product on matrices.

Thus H is a pertubation of the identity matrix, namely $H = B \times \mathrm{id} + A$, where

$$A = C^2 \theta \otimes \theta + C\alpha \otimes \theta + C\theta \otimes \alpha.$$

Since θ projects onto the direction e, $\theta(v) = 0$ for any vector v perpendicular to e. Similarly, it can be seen that

$$\begin{aligned} A(u \otimes v + v \otimes u) &= C^2(\theta u) \otimes (\theta v) + C^2(\theta v) \otimes (\theta u) \\ &\quad + C(\alpha u) \otimes (\theta v) + C(\alpha v) \otimes (\theta u) \\ &\quad + C(\theta u) \otimes (\alpha v) + C(\theta v) \otimes (\alpha u) \\ &= 0 \text{ for all } u, v \text{ perpendicular to } e. \end{aligned}$$

On the other hand, A is non-zero (in general) on vectors of the form $v \otimes e + e \otimes v$ for $v \in V$. This suggests that we define the subspace

$$U = \mathrm{span}\{w \otimes e + e \otimes w : w \in V\}.$$

It is easy to verify that

$$U^{\perp} = \mathrm{span}\{u \otimes v + v \otimes u : u, v \in e^{\perp}\}.$$

We have already shown that $U^{\perp} \subset \ker A$ and, since A is symmetric, we deduce that $\mathrm{im}\, A \subset U$, which is also evident by direct calculation. Thus A has the decomposition

$$A = \begin{array}{cc} \begin{array}{cc} \;U\; & \;U^{\perp} \end{array} & \\ \left[\begin{array}{c|c} * & 0 \\ \hline 0 & 0 \end{array}\right] & \begin{array}{l} U \\ U^{\perp} \end{array} \end{array}$$

The $*$ symbol represents the non-zero part of A. At this stage it is apparent that away from U, H is trivial, so the inversion problem reduces to the problem of inverting H on the n-dimensional subspace U.

However, up to this point we have ignored a crucial point: the random variables d_{ij} satisfy

$$d_{ii} \equiv 0 \text{ for } i = 1, \dots, n.$$

(In other words the distance of a gene from itself is zero.) Instead of working on the space W, we are really dealing with the restriction of H to the space

$$\hat{W} = \mathrm{span}\{\frac{1}{2} e_i \otimes e_j + \frac{1}{2} e_j \otimes e_i : i < j\}.$$

Note that the inequality in the indices is strict here. If we define $P : W \to W$ by

$$P(\frac{1}{2} e_i \otimes e_j + \frac{1}{2} e_j \otimes e_i) = \begin{cases} \frac{1}{2} e_i \otimes e_j + \frac{1}{2} e_j \otimes e_i & \text{if } i \neq j \\ 0 & \text{when } i = j \end{cases}$$

then $\hat{W} = P(W)$ and the restriction of H to $\hat{W}$, denoted $\hat{H}$, is given by

$$\begin{aligned}\hat{H} &= PHP \\ &= PAP + P\,(B \times \mathrm{id})\,P.\end{aligned}$$

It is really the map $\hat{H}$ that we need to invert; $P\,(B \times \mathrm{id})\,P$ is simply a multiple of the identity on $\hat{W}$..

Essentially we want to identify a decomposition for $\hat{H}$ equivalent to the one above. Let $\hat{U} = P(U)$ and let $\hat{U}^{\perp}$ be the orthogonal complement of $\hat{U}$ in $\hat{W}$, so that

$$\hat{W} = \hat{U} \oplus \hat{U}^{\perp}$$

As we showed above, im $A \subset U$, so

$$\text{im } PAP \subset P(U) = \hat{U}.$$

Since A and P are symmetric, we therefore have the desired decomposition:

$$PAP = \begin{array}{c|c|c} \hat{U} & \hat{U}^{\perp} & \\ \hline * & 0 & \hat{U} \\ \hline 0 & 0 & \hat{U}^{\perp} \end{array}$$

It follows that

$$\hat{H}^{-1}(x) = \left(P(A+B)P|_{\hat{U}}\right)^{-1}(x_{\hat{U}}) + B^{-1}(x_{\hat{U}^{\perp}}) \tag{8}$$

where

$$x = x_{\hat{U}} \oplus x_{\hat{U}^{\perp}} \tag{9}$$

represents the $\hat{U}$ and $\hat{U}^{\perp}$ decomposition of $x \in \hat{W}$. The first term of (8) denotes the restriction of $P(A+B)P$ to the n-dimensional space $\hat{U}$. The inversion problem therefore reduces to the question of inverting this component of the map.

To address this problem it is useful to work with an orthonormal basis of $\hat{U}$. It can be shown that

$$w_k = \phi(\frac{1}{2}e_k \otimes e + \frac{1}{2}e \otimes e_k - e_k \otimes e_k) + \psi(e \otimes e - \sum_l e_l \otimes e_l)$$

for $k = 1, \ldots, n$ defines an orthonormal basis of $\hat{U}$ when

$$\phi = \left(\frac{2}{n-2}\right)^{\frac{1}{2}}$$

and

$$\psi = -\frac{1}{n}\left(\left(\frac{2}{n-2}\right)^{\frac{1}{2}} + \left(\frac{1}{n-1}\right)^{\frac{1}{2}}\right).$$

In this basis $\hat{H}$ is represented by the matrix

$$\begin{aligned}\Lambda_{ij} &= B\delta_{ij} + \langle w_i, PAP(w_j)\rangle \\ &= \left(C(n-2)a_i + B\right)\delta_{ij} + \frac{1}{2}C\phi(n-2)\left(\phi + 2(n-1)\psi\right)(a_i + a_j) \\ &\quad + \frac{1}{2}C\left(\phi + 2(n-1)\psi\right)^2\left(\sum_l a_l\right) + C^2(n-1)^2(\phi + n\psi)^2.\end{aligned}$$

The matrix defines a linear map $\Lambda : V \to V$

$$\Lambda = M + \omega e^T + e\omega^T$$

where

$$M = \operatorname{diag}\left(C(n-2)a_k + B : k = 1, \dots, n\right),$$

and $\omega \in V$ is defined by

$$\begin{aligned}\omega = {} & \frac{1}{2}C\phi(n-2)\left(\phi + 2(n-1)\psi\right)a \\ & + \left(\frac{1}{2}C^2(n-1)^2(\phi+n\psi)^2 + \frac{1}{4}C\left(\phi + 2(n-1)\psi\right)^2 \sum_l a_l\right)e. \qquad (10)\end{aligned}$$

The matrix Λ is a rank-2 perturbation of M and its inverse is given by the Sherman–Morrison–Woodbury formula [6]:

$$\Lambda^{-1} = M^{-1} + \frac{c_{\omega\omega}\hat{e}\hat{e}^T + c_{ee}\hat{\omega}\hat{\omega}^T - (1 + c_{\omega e})(\hat{\omega}\hat{e}^T + \hat{e}\hat{\omega}^T)}{(1 + c_{\omega e})^2 - c_{\omega\omega}c_{ee}} \qquad (11)$$

where

$$\hat{e} = M^{-1}e, \quad \hat{\omega} = M^{-1}\omega$$

and

$$c_{\omega\omega} = \langle M^{-1}\omega, \omega\rangle, \quad c_{ee} = \langle M^{-1}e, e\rangle, \text{ and } \quad c_{\omega e} = \langle M^{-1}\omega, e\rangle.$$

We now have all the elements in place for a complete algebraic inverse to the map $\hat{H}$ defined by the covariance tensor (1). The next section puts these elements together and specifies the inversion algorithm.

4. The inversion algorithm

Suppose that we want to compute $\hat{H}^{-1}x$ for some symmetric matrix x_{ij} that is zero on the diagonal. The inversion algorithm has the following steps.

(1) On account of the decomposition (8), the $\hat{U}$ and $\hat{U}^\perp$ components of x can be dealt with independently. The $\hat{U}$ component of x is be obtained by taking the inner product of x with the vectors w_k that form our basis of $\hat{U}$. We therefore define the vector $\xi \in V$ by

$$\xi_k = \langle x, w_k\rangle = \frac{1}{2}\phi\left(\sum_i x_{ik} + \sum_j x_{kj}\right) + \psi\sum_{ij} x_{ij}. \qquad (12)$$

(2) The inverse of $\hat{H}$ on $\hat{U}$ is given by Equation (11). Let η be the inverse of ξ:

$$\eta = \Lambda^{-1}\xi.$$

(3) The final step is to combine this result with the $\hat{U}^\perp$ component of the inverse. Using the decompositions (8) and (9), and since $B^{-1}(x_{\hat{U}^\perp}) = B^{-1}(x) - B^{-1}(x_{\hat{U}})$ it follows that

$$\hat{H}^{-1}(x) = \left(P(A+B)P|_{\hat{U}}\right)^{-1}(x_{\hat{U}}) - B^{-1}(x_{\hat{U}}) + B^{-1}(x) \qquad (13)$$

The first term of (13) is given using the result of step 2:

$$\begin{aligned}\left(P(A+B)P|_{\hat{U}}\right)^{-1}(x_{\hat{U}}) &= \sum_{k=1}^{n} \eta_k w_k \\ &= \frac{\phi}{2}P(\eta \otimes e + e \otimes \eta) + \psi\langle e, \eta\rangle P(e \otimes e).\end{aligned}$$

By replacing η with $(\eta - B^{-1}\xi)$ in this equation we obtain the first two terms of (13). The full inverse is then given by

$$\hat{H}^{-1}(x) = B^{-1}x + \frac{\phi}{2}P\left((\eta - B^{-1}\xi)\otimes e + e\otimes(\eta - B^{-1}\xi)\right) + \psi\langle e, \eta - B^{-1}\xi\rangle P(e\otimes e).$$

The computational complexity of the algorithm can be obtained by analysing each step. It is easy to see that overall there are $O(n^2)$ multiplications and $O(n^2)$ additions. Moreover, in our application we actually use the inverse covariance tensor to evaluate inner products like $x\cdot\hat{H}^{-1}x$ where x is a symmetric matrix with zero diagonal and the dot product is the standard inner product between matrices. If the partial sums in Equation (12) are already known for the matrix x, then the inner product $x\cdot\hat{H}^{-1}x$ can be evaluated in $O(n)$ operations.

5. Discussion

The crucial advantage of the probabilistic model chosen in this paper is that it generates covariance matrices which are mild perturbations of the identity, so ensuring inexpensive inversion. However, the basis chosen is highly reminiscent of of similar bases chosen to elucidate the structure of Euclidean distance matrices, and we mention this interesting connection. We recall that an $n\times n$ matrix A is a Euclidean distance matrix if there exist vectors $u_1,\ldots,u_n\in\mathbb{R}^n$ for which

$$A_{ij} = \|u_i - u_j\|^2, \qquad 1\le i,j\le n,$$

where $\|\cdot\|$ denotes the Euclidean norm. Such matrices were characterized by I. J. Schoenberg [8], who proved that a symmetric matrix M, whose diagonal elements vanish, is a Euclidean distance matrix if and only if $v^T M v \le 0$ when v is orthogonal to e, the vector of all ones defined at the beginning of Section 3. The theory of such matrices is highly relevant to the linear of radial basis functions and learning theory; see, for example, [1].

References

1. B. J. C. Baxter, *Conditionally positive functions and p-norm distance matrices*, Constr. Approx. **7** (1991), 427–440.
2. W. J. Bruno, N. D. Socci, and A. L. Halpern, *Weighted neighbour-joining: A likelihood-based approach to distance-based phylogeny reconstruction*, Molecular Biology and Evolution **17** (2000), 189–197.
3. L. L. Cavalli-Sforza and A. W. F. Edwards, *Phylogenetic analysis: Models and estimation procedures*, American Journal of Human Genetics **19** (1967), 233–257.
4. J. Felsenstein, *Inferring phylogenies*, Sinauer, 2004.
5. O. Gascuel, *Bionj: An improved version of the neighbour-joining algorithm based on a simple model of sequence data*, Molecular Biology and Evolution **14** (1997), 685–695.
6. G. H. Golub and C. F. Van Loan, *Matrix computations*, 3 ed., Johns Hopkins University Press, 1996.
7. N. Satou and M. Nei, *The neighbour-joining method: A new method for reconstructing phylogenetic trees*, Molecular Biology and Evolution **4** (1987), 406–425.
8. I. J. Schoenberg, *Remarks to Maurice Fréchet's article 'Sur la définition axiomatique d'une classe d'espace distanciés vectoriellemetn applicable sur l'espace d'Hilbert'*, Ann. of Math. **36** (1935), 724–736.