A peer-reviewed version of this preprint was published in PeerJ on 25 September 2014.

View the peer-reviewed version (peerj.com/articles/583), which is the preferred citable publication unless you specifically need to cite this preprint.

Gil M. 2014. Fast and accurate estimation of the covariance between pairwise maximum likelihood distances. PeerJ 2:e583 https://doi.org/10.7717/peerj.583

Fast and Accurate Estimation of the Covariance between Pairwise Maximum Likelihood Distances

Manuel Gil[1]

[1]**Institute of Molecular Life Sciences,University of Zurich, Zurich, Switzerland. Swiss Institute of Bioinformatics, Lausanne, Switzerland.**

ABSTRACT

Pairwise evolutionary distances are a model-based summary statistic for a set of molecular sequences. They represent the leaf-to-leaf path lengths of the underlying phylogenetic tree. Estimates of pairwise distances with overlapping paths covary because of shared mutation events. It is desirable to take these covariance structure into account in any process that compares or combines distances to increase precision. In this paper, we present a fast estimator for the covariance of two pairwise maximum likelihood distances, estimated under general Markov models. The estimator is based on a conjecture (going back to Nei and Jin, 1989) which links the covariance to path lengths. We prove it here under a simple symmetric substitution model. In a simulation, we show that our estimator outperforms previously published ones in terms of the mean squared error.

Keywords: Pairwise distance, Maximum likelihood, Correlation, Covariance, Alignment

INTRODUCTION

Phylogenetic trees are one of the most important representations of the evolutionary relationship between homologous genomic sequences. Their relatedness can be summarized by a set of pairwise evolutionary distances representing the leaf-to-leaf path lengths of the underlying tree. Such distances are usually estimated by maximum likelihood (ML) assuming a Markovian model of character substitution (Yang, 2006).

Besides substitutions, a process of insertions and deletions of sequence fragments plays a major role in the evolution of molecular sequences. As a consequence, homologous characters –i.e. the ones related by substitutions only– have to be identified prior to distance estimation. A consistent hypothesis of character homology is provided by multiple sequence alignments (MSAs). Alternatively, the sequences can be aligned pairwise, for instance, by dynamic programming to obtain optimal pairwise alignments (OPAs) (Needleman and Wunsch, 1970).

Pairwise distance methods are generally faster and also simpler than likelihood based approaches that operate directly on sequence data. For that reason, they have often been chosen as an input to large-scale genomic and phylogenetic analyses. Further, distance tree methods are used to produce starting trees for ML tree estimation from MSAs (Guindon et al., 2010; Stamatakis, 2014; Vinh and von Haeseler, 2004; Gil et al., 2013) and guide trees in progressive MSA methods (e.g. Löytynoja and Goldman, 2008; Katoh et al., 2005).

The speed benefits may affect accuracy due to a potential loss of information involved in the reduction of the sequence data (Steel et al., 1988). However, this idea has recently been challenged in the context of tree estimation (Roch, 2010). Roch proposed to take advantage of higher order information using the correlations among the pairwise distances, which result from common mutation events on shared paths of the underlying tree. Indeed, most current practical distance tree methods assume statistical independence and do not account for distance covariance (Mihaescu and Pachter, 2008). The BioNJ algorithm, which uses a first-order model of covariance, is a notable exception (Gascuel, 1997). Generally, any process that compares or combines distances profits from a higher precision when the covariance structure is taken into account.

Estimators for the covariance between pairwise ML distance estimates have been proposed for certain

mechanistic substitution models (Tajima and Nei, 1984; Nei and Jin, 1989; Bulmer, 1991) and for general Markov models by Susko (2003). Susko's estimator requires an MSA and has a linear time complexity in the sequence length. We have previously derived an adaptation for OPAs with similar complexity. In this paper, we present a constant time estimator for general Markov models, applicable to OPAs and MSAs. To this end, we prove conjecture (from Nei and Jin, 1989; Bulmer, 1991) for a simple symmetric substitution model and extend it to general models. In a simulation we evaluate our estimator and compare it to previously published ones.

METHODS

Preliminaries

Denote $A = \{x_i\}_{i=1}^n$ a pairwise alignment consisting of n homologous i.i.d. character-pairs x_i (e.g. nucleotides, amino acids, or codons, but no insertion-deletions). The likelihood of having the two sequences in A separated by an evolutionary distance d is (Felsenstein, 1981)

$$L(A|d) = \prod_{i=1}^{n} p(x_i, d), \tag{1}$$

where $p(x_i, d)$ is the probability of the character-pair x_i at d. The ML estimator of the true distance d_t is

$$\hat{d} = \arg\max_{d \geq 0} L(A|d). \tag{2}$$

While for simple mechanistic substitution models the maximum can be expressed analytically, it is usually found numerically for empirical and complex mechanistic models using the Newton-Raphson method. Let $I_n(d)$ denote the Fisher information for d, i.e.

$$I_n(d) = -nE\left[\frac{\partial^2}{\partial d^2} \log p(X, d)\right], \tag{3}$$

where X is a random variable of which the x_i are realizations. The asymptotic variance of $\hat{d}$ is provided by standard theory (e.g. Pawitan, 2001)

$$V(\hat{d}) = I_n(d_t)^{-1}. \tag{4}$$

It can be estimated by evaluating the inverse of the Fisher information at $\hat{d}$ (hereafter *ML variance*)

$$\hat{V}(\hat{d}) = I_n(\hat{d})^{-1}, \tag{5}$$

or, alternatively, from A and $\hat{d}$ by a sample average:

$$\hat{V}_A(\hat{d}) = -\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial d^2} \log p(x_i, d)\Big|_{d=\hat{d}}\right)^{-1} = -\left(\sum_{i=1}^{n} \frac{\partial^2}{\partial d^2} \log p(x_i, d)\Big|_{d=\hat{d}}\right)^{-1}. \tag{6}$$

We distinguish three topological relations relevant for covariance estimation between any two pairwise distance estimates (Dessimoz and Gil, 2008). First, the relation *dependence*, where two distances share some common evolution (e.g. d_{ij} and d_{kl} in Figure 1). Second, the similar relation *triplet*, where two distances additionally share a sequence (e.g. d_{ij} and d_{kj}). Third, the case *independence*, where the distances are independent (e.g. d_{ik} and d_{jl}). Note that the second case can be conceptually reduced to the first (with $j = l$ and $\delta_j = \delta_l = 0$). Further, we assume that mutation events on different edges of a tree are independent, thus, the distances in the independence-case have zero covariance. Therefore, our derivations will focus on the dependence-case without loss of generality.

General ML theory provides covariance estimates if all unknown parameters are estimated jointly. For instance, in ML tree reconstruction, the variance/covariance matrix can be estimated from the observed Fisher information matrix. However, the estimation of the pairwise distances considered here is done separately for each distance so that general ML theory can not be applied. Susko (2003) has derived an interesting estimator for the covariance of two distances $\hat{d}_i$ and $\hat{d}_j$ estimated from pairwise alignments

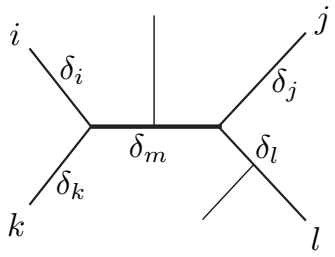

Figure 1. Unrooted tree relating six sequences. The labeled sequences $\{i, j, k, l\}$ define a subtree of four sequences, a quartet. The δ's indicate the branch lengths of the quartet, representing number of character changes. Under Markovian substitution models the ML estimates of the pairwise distances $d_{ij} = \delta_i + \delta_m + \delta_j$ and $d_{kl} = \delta_k + \delta_m + \delta_l$ covary because of the δ_m common mutation events.

which are induced by an MSA. It is a sample average based on the following expression (hereafter *Susko-covariance*)

$$\operatorname{cov}(\hat{d}_i, \hat{d}_j) = nV_iV_jE\left[\frac{\partial}{\partial d}\log p_i(X,d)\Big|_{d=d_{t,i}} \cdot \frac{\partial}{\partial d}\log p_j(X,d)\Big|_{d=d_{t,j}}\right]. \tag{7}$$

Here, $d_{t,i}$ and $d_{t,j}$ are the true distances, the random-variable X stands for quartets of homologous characters (as opposed to pairs in Equation 3), and $p_i(X,d)$ denotes the probability for the i-th pair in X at the distance d. The Susko-covariance has two limitation. First, it requires an MSA, i.e. it is no applicable to distances derived from OPAs (for a discussion see Dessimoz and Gil, 2008). Second, the $O(n)$ computation time may become prohibitive in large scale studies. Alternatively, a nonparametric bootstrap can be used (Efron and Tibshirani, 1993), but it takes substantially longer computation times and an requires an MSA too.

We have previously presented two estimators to tackle the limitations. They work with both MSAs and OPAs. The first method is a numerical approximation to the variance of the difference between two distances involving a common sequence. It runs in constant time with respect to the sequence-length (Dessimoz et al., 2006). This leads to a fast covariance estimator for the triplet-case (hereafter referred to as *triplet-covariance*). The second method is based on Susko's theory and shares the linear time complexity (Dessimoz and Gil, 2008, hereafter *anchor-covariance*). It was specifically designed to bypass the problem of inconsistent homology inference between OPAs using the concept of *anchors* –a globally consistent subset of aligned character pairs. In this paper, we propose a fast and general approach we term *branch-covariance*. It is motivated by an analytic result obtained under the simple r-state symmetric model.

r-state symmetric model

To obtain analytic results we will work with the r-state symmetric model, also know as the N_r model (Neyman, 1971). It is a generalization of the Jukes-Cantor model (Jukes and Cantor, 1969), which has four states ($r = 4$), to r character-states. The N_r model assumes a uniform distribution of states at the root, and equal rates of transitions between any two distinct character states. The probability to observe a mutation after time t is

$$p_m(t) = \beta\left(1 - \mathrm{e}^{-\frac{\alpha t}{\beta}}\right), \quad \beta = \frac{r-1}{r}, \tag{8}$$

where α is the total rate of substitution. Thus, if two sequences are separated by t, the distance between them will be $d = \alpha t$.

Because of the symmetries in the model, the number of differing sites I in a given pairwise alignment of length n is a sufficient statistics for the pairwise ML distance

$$\hat{d} = -\beta \ln\left(1 - \frac{I}{n\beta}\right). \tag{9}$$

An estimator for the variance of $\hat{d}$ can be obtained by applying Equation 4 derived from the likelihood function. Alternatively, for models estimating distances from proportions of differing sites, the variance

can be approximated by the Delta technique. This has been done by Kimura and Ohta (1972) for $r = 4$ and generalized by Tajima and Nei (1984) to

$$\hat{V}(\hat{d}) = \beta \left[(1-\beta)e^{\frac{2d}{\beta}} + (2\beta - 1)e^{\frac{d}{\beta}} - \beta \right] / n. \tag{10}$$

We are going to use the Delta technique to derive an estimator for the covariance of two N_r ML distances $d_{ij} = \delta_i + \delta_m + \delta_j$ and $d_{kl} = \delta_k + \delta_m + \delta_l$ in the dependence-case (Figure 1). Nei and Jin (1989) used an informal argument to propose the following expression with $\beta = 3/4$:

$$\mathrm{cov}(\hat{d}_{ij}, \hat{d}_{kl}) = \beta \left[(1-\beta)e^{\frac{2\delta_m}{\beta}} + (2\beta - 1)e^{\frac{\delta_m}{\beta}} - \beta \right] / n. \tag{11}$$

The equation originates from the assumption, that the covariance of two distance estimates with an underlying shared path length δ_m, is formally equivalent to the variance (Equation 10) of an ML estimate of a pairwise distance δ_m. Indeed, Bulmer (1991) presented a proof for $\beta = 3/4$ for the triplet-case and conjectured that Equation 11 with $j = l$ and $\delta_j = \delta_l = 0$ is true for any β. To the best of our knowledge Equation 11 has not been proven yet in its most general form, i.e. for the dependence-case and any β.

We will first compute

$$\mathrm{cov}(I_{ij}, I_{kl}) = n\,\mathrm{cov}(S_{ij}, S_{kl}) = nE\left[S_{ij}S_{kl}\right] - nE[S_{ij}]E[S_{kl}], \tag{12}$$

where S_{ij} is a random variable indicating whether sequences i and j are identical ($S_{ij} = 0$) or different ($S_{ij} = 1$) at a particular site. Subsequently, we will apply the Delta method to obtain Equation 11 from $\mathrm{cov}(I_{ij}, I_{kl})$. We start by noting that

$$E[S_{ij}] = \Pr(S_{ij} = 1) = p_m(d_{ij}). \tag{13}$$

Therefore, the problem reduces to computing

$$E[S_{ij}S_{kl}] = \Pr(S_{ij} = 1 \wedge S_{kl} = 1). \tag{14}$$

In the following we will represent the quartet from Figure 1 by the symbol  , where the terminal node of the upper left branch corresponds to i. Furthermore, we are going to mark a branch with $\circ$ if a particular site in the evolving sequence is in a different state at the endpoints of the branch. As an example, we look at the pattern  . Here, the site in question changed its state on the branches leading to $\{j,k,l\}$ but did not change its state on the branch leading to i and on the middle branch. A particular labeling of the nodes with characters from the alphabet of N_r for the given pattern has probability

$$\Pr(\text{  }) = \frac{1}{r} \cdot (1 - p_m(\delta_i)) \cdot \frac{p_m(\delta_j)}{(r-1)} \cdot \frac{p_m(\delta_k)}{(r-1)} \cdot \frac{p_m(\delta_l)}{(r-1)} \cdot (1 - p_m(\delta_m)). \tag{15}$$

For this pattern, there are $r(r-1)^3$ possible labellings, of which only $r(r-1)^2(r-2)$ satisfy $S_{ij} = 1 \wedge S_{kl} = 1$. Symmetrical mutation patterns, like for example  have the same number of labelings leading to $S_{ij} = 1 \wedge S_{kl} = 1$ but different mutation probabilities. We consider now all the patterns (grouped by symmetry) and corresponding labelings for the desired condition:

$$\begin{aligned}
&\Pr(S_{ij} = 1 \wedge S_{kl} = 1) = \\
&\quad \Pr(\text{  }) \cdot r(r-1) \\
&+ \left[\Pr(\text{  }) + \Pr(\text{}) + \Pr(\text{  }) + \Pr(\text{  }) \right] \cdot r(r-1)^2 \\
&+ \left[\Pr(\text{}) + \Pr(\text{  }) + \Pr(\text{}) + \Pr(\text{  }) \right] \cdot r(r-1)(r-2) \\
&+ \left[\Pr(\text{}) + \Pr(\text{}) + \Pr(\text{}) + \Pr(\text{}) \right] \cdot r(r-1)^2(r-2) \\
&+ \left[\Pr(\text{}) + \Pr(\text{}) + \Pr(\text{}) + \Pr(\text{}) \right] \cdot r(r-1)(r-2)^2
\end{aligned} \tag{16}$$

$$+ \Pr(\text{}) \cdot r(r-1)^2(r-2)^2$$

$$+ \left[\Pr(\text{}) + \Pr(\text{})\right] \cdot r(r-1)\left[(r-1)+(r-2)^2\right]$$

$$+ \left[\Pr(\text{}) + \Pr(\text{}) + \Pr(\text{}) + \Pr(\text{})\right] \cdot r(r-1)\left[(r-1)+(r-2)^2\right](r-2)$$

$$+ \Pr(\text{}) \cdot r(r-1)\left[(r-1)+(r-2)^2\right]^2.$$

Using Maple (script in Appendix A.1) we find that the expression simplifies to

$$E[S_{ij}S_{kl}] = \beta\left(1-\beta e^{-\frac{d_{ij}}{\beta}} - \beta e^{-\frac{d_{kl}}{\beta}} + (1-\beta)e^{-\frac{d_{ij}+d_{kl}-2\delta_m}{\beta}} + (2\beta-1)e^{-\frac{d_{ij}+d_{kl}-\delta_m}{\beta}}\right). \tag{17}$$

Plugging Equations 17 and 13 in 12 we obtain

$$\operatorname{cov}(I_{ij},I_{kl}) = n\beta\left((1-\beta)e^{-\frac{d_{ij}+d_{kl}-2\delta_m}{\beta}} + (2\beta-1)e^{-\frac{d_{ij}+d_{kl}-\delta_m}{\beta}} - \beta e^{-\frac{d_{ij}+d_{kl}}{\beta}}\right). \tag{18}$$

We turn to the Delta method. The function $\hat{d}_{ij}(I_{ij})$ can be approximated by a first-order Taylor series around $E[I_{ij}] = np_m(d_{ij})$

$$\hat{d}^{\star}_{ij}(I_{ij}) = \hat{d}_{ij}(E[I_{ij}]) + \hat{d}'_{ij}(E[I_{ij}])(I_{ij}-E[I_{ij}]), \tag{19}$$

where

$$\hat{d}'_{ij}(E[I_{ij}]) = \left.\frac{\partial \hat{d}_{ij}(I_{ij})}{\partial I_{ij}}\right|_{I_{ij}=E[I_{ij}]} = \left.\left(n-\frac{I_{ij}}{\beta}\right)\right|_{I_{ij}=E[I_{ij}]} = \frac{e^{\frac{d_{ij}}{\beta}}}{n}. \tag{20}$$

The covariance of $\hat{d}_{ij}$ and $\hat{d}_{kl}$ is asymptotically equal to the covariance of $\hat{d}^{\star}_{ij}$ and $\hat{d}^{\star}_{kl}$

$$\operatorname{cov}\left(\hat{d}_{ij}(I_{ij}),\hat{d}_{kl}(I_{kl})\right) \sim \operatorname{cov}\left(\hat{d}^{\star}_{ij}(I_{ij}),\hat{d}^{\star}_{kl}(I_{kl}))\right) \tag{21}$$

$$= \hat{d}'_{ij}(E[I_{ij}])\hat{d}'_{kl}(E[I_{kl}])\operatorname{cov}(I_{ij},I_{kl}) \tag{22}$$

$$= \beta\left[(1-\beta)e^{\frac{2\delta_m}{\beta}} + (2\beta-1)e^{\frac{\delta_m}{\beta}} - \beta\right]/n. \qquad \square \tag{23}$$

We provide a Maple program implementing all steps of the proof in Appendix A.1.

Covariance under general Markov models

We have shown under N_r that the covariance of two distance estimates with an underlying shared path length δ_m is asymptotically equal to the variance of an ML estimate of a true pairwise distance δ_m, i.e.

$$\operatorname{cov}(\hat{d}_{ij},\hat{d}_{kl}) \sim V(\hat{d}|d_t = \delta_m). \tag{24}$$

We conjecture that the relationship holds for general substitution models and apply it to derive a covariance estimator. To this end, we first discuss how the covariance can be written in terms of a rate matrix Q, an equilibrium frequency vector π (which together fully specify a substitution model), and δ_m. In the next section, we show how δ_m can be estimated from the input distances by the method of weighted least squares (WLS).

According to our conjecture and Equation 4 we express the desired covariance by

$$\operatorname{cov}(\hat{d}_{ij},\hat{d}_{kl}) = -\frac{1}{n}E\left[\left.\frac{\partial^2}{\partial d^2}\log p(X,d)\right|_{d=\delta_m}\right]^{-1}. \tag{25}$$

The expected value expression can be written as

$$\sum_{\forall(u,v)}\left[p(x_{uv},d)\frac{\partial^2}{\partial d^2}\log p(x_{uv},d)\right]_{d=\delta_m} = \sum_{\forall(u,v)}\left[\frac{\partial^2}{\partial d^2}p(x_{uv},d) - \frac{\left(\frac{\partial}{\partial d}p(x_{uv},d)\right)^2}{p(x_{uv},d)}\right]_{d=\delta_m}, \tag{26}$$

where the summation goes over all possible character pairs. In terms of a rate matrix Q and an equilibrium frequency vector π the probability of a character pair ($k = 0$) and the derivatives ($k > 0$) of the probability with respect to the distance are

$$\frac{\partial^{(k)}}{\partial d^{(k)}} p(x_{uv}, d) = \pi_u \left[Q^k e^{Qd} \right]_{uv}. \tag{27}$$

Plugging Equations 26 and 27 in 25 we obtain

$$\mathrm{cov}(\hat{d}_{ij}, \hat{d}_{kl}) = -\frac{1}{n} \left[\sum_{\forall (u,v)} \pi_u \left(\left[Q^2 e^{Q\delta_m} \right]_{uv} - \frac{\left[Q e^{Q\delta_m} \right]^2_{uv}}{\left[e^{Q\delta_m} \right]_{uv}} \right) \right]^{-1}. \tag{28}$$

To save computation time, we can discretise the distance space in the relevant range to some desired level of accuracy, precompute the expected value expressions and store them in a hash table. A covariance estimator is then obtained by substituting δ_m in Equation 28 by its WLS estimate, which we derive in the next section.

Topological relation and path length

Equation 28 expresses the covariance in the dependence case as a function of the shared path length δ_m. To obtain a covariance estimator, we determine first whether the two distances in question are dependent and, in case they are, estimate δ_m. We will do that by WLS using the six pairwise distance estimates $\{\hat{d}_{uv}\}$ between the four sequences i, j, k, l and their variances $\{v_{uv}\}$. The sequences can be related by three topological configurations:

$$T_1 : ((i,k),(j,l)), \quad T_2 : ((i,l),(j,k)), \quad T_3 : ((i,j),(k,l)),$$

where T_1, T_2 map to the dependence case and T_3 corresponds to the independence case. An argument set out in Appendix A.2 shows that the weighted sum of squares (S) for each of the topologies can be expressed in a simple form which is fast to compute:

$$S(T_1) = \frac{\hat{d}_{ij} + \hat{d}_{kl} - \hat{d}_{il} - \hat{d}_{jk}}{v_{ij} + v_{kl} + v_{il} + v_{jk}}, \quad S(T_2) = \frac{\hat{d}_{ij} + \hat{d}_{kl} - \hat{d}_{ik} - \hat{d}_{jl}}{v_{ij} + v_{kl} + v_{ik} + v_{jl}}, \quad S(T_3) = \frac{\hat{d}_{ik} + \hat{d}_{jl} - \hat{d}_{il} - \hat{d}_{jk}}{v_{ik} + v_{jl} + v_{il} + v_{jk}}. \tag{29}$$

The best fitting topology is determined by $\mathrm{argmin}_{T_i}\{S(T_i)\}$. If this results in T_3 the desired covariance is zero, otherwise we need to estimate δ_m. The WLS estimates are (see Appendix A.2):

$$2\hat{\delta}_m(T_1) = \frac{(\hat{d}_{ij} + \hat{d}_{kl})(v_{il} + v_{jk}) + (\hat{d}_{il} + \hat{d}_{jk})(v_{ij} + v_{kl})}{v_{ij} + v_{kl} + v_{il} + v_{jk}} - (\hat{d}_{ik} + \hat{d}_{jl}), \tag{30}$$

$$2\hat{\delta}_m(T_2) = \frac{(\hat{d}_{ij} + \hat{d}_{kl})(v_{ik} + v_{jl}) + (\hat{d}_{ik} + \hat{d}_{jl})(v_{ij} + v_{kl})}{v_{ij} + v_{kl} + v_{ik} + v_{jl}} - (\hat{d}_{il} + \hat{d}_{jk}). \tag{31}$$

Since these are the estimators for unconstrained WLS they can result in negative values, in which case we estimate the covariance to be zero.

Simulation settings

To evaluated the performance of the various covariance estimators we adopted the same simulation approach as in one of our previous studies (Dessimoz and Gil, 2008). We sampled 100 random quartets from a tree of life on 352 species. The tree was inferred by the *LeastSquaresTree* function (Gil and Gonnet, 2009) included in the Darwin package (Gonnet et al., 2000) using pairwise distance and variance data from the OMA project (Dessimoz et al., 2005). We applied a uniformly distributed $\mathrm{U}(0.5, 2)$ expansion/contraction factor on each quartet to also explore extremer regions of the branch-length space, while preserving the relative branch-length structure of the original tree.

For each dilated model quarted we generated $10,000$ times three random amino-acid sequences of length $m = \{200, 500, 800\}$ and mutated them along the quartet assuming the GCB substitution model (Gonnet et al., 1992). The entire simulation procedure was run twice, once without any insertion-deletions to produce ungapped alignments to test the methods under the true models (i.e. without the effect of

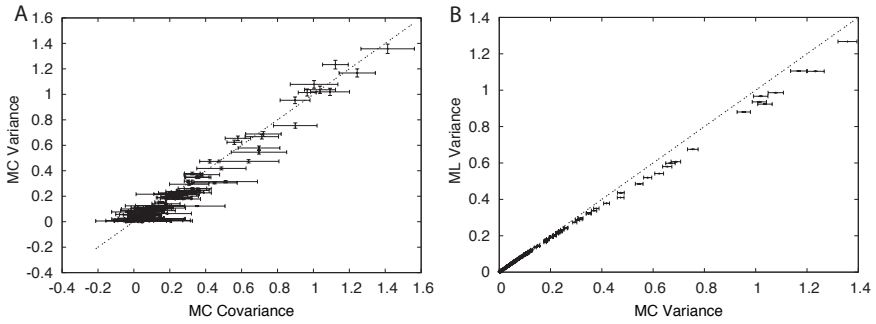

Figure 2. Components of branch-covariance for sequence-length 10,000. Error-bars indicate 95% confidence intervals. **A.** Monte Carlo covariance versus Monte Carlo variance shows that the relationship derived under the N_r model also holds for the empirical substitution model tested here. **B.** Monte Carlo variance versus mean of the ML variance shows that the ML variance is negatively biased.

alignment errors), and once introducing gaps of Zipfian distributed length (Benner et al., 1993). The gapped sequences were aligned by global pairwise dynamic programming with the *Align* function from Darwin to obtain OPAs. ML pairwise distances were estimated with the *EstimatePam* function. The sample variance and covariance over the 10,000 samples (hereafter *Monte Carlo-variance* and *Monte Carlo-covariance*) served as a reference values, as they are unbiased estimators of the true values.

To test our asymptotic conjecture (Equation 24) we repeated the ungapped simulation with very long sequences (10,000 amino-acids). Then, we extracted the length δ_m of the middle branch of each model quartet, generated 10,000 random sequences, and mutated each with a distance δ_m according to the GCB model. Finally, we estimated the ML distance and the ML variance between each resulting pair of sequences, and computed the Monte Carlo-variance.

RESULTS AND DISCUSSION

Evaluation of basic components of branch-covariance

We have tested the validity of Equation 24, which was derived under the N_r model, and the accuracy of the ML variance (Equation 5). A plot of the Monte Carlo-variance versus the Monte Carlo-covariance for long sequences (10,000 amino-acids) corroborates the conjecture under the GCB model and suggests that the result is valid in general (Figure 2A). The branch-covariance relies on the ML variance to approximate the true variance. A comparison with the Monte Carlo-Variance shows that it underestimates the variance for large samples and with correct alignments (Figure 2B). Therefore, we expected the branch-covariance to inherit the negative bias.

Evaluation of estimators

We present now the performance of the branch-covariance under the true model and compare it with the Susko- and triplet-covariance (Figure 3). To this end, we discuss the tree topological cases –dependence, independence, and triplet– separately.

In the dependence case the Susko-covariance is unbiased. The branch-covariance lies in most cases within the 95% confidence interval of the Monte Carlo covariance; when it lies outside then it underestimates. The negative bias confirms the prediction from previous section. In the independence case, where the true covariance is zero, both estimators have a positive bias of comparable magnitude, though the branch-covariance appears to have a lower bias with increasing sequence length. For the anchor-covariance the negative bias is no surprise; it returns by construction non-negative values. For the triplets, as in the dependence case, Susko's estimator appears to be unbiased and the branch-covariance shows a minor negative bias. The triplet-covariance has a positive bias of comparable magnitude to the one of the branch-covariance. Although the lack of bias is an attractive feature of an estimator, it does not guarantee a low total error. Therefore, it is instructive to look at the mean square error (MSE). Indeed, the branch-covariance has a lower MSE than the Susko-covariance under all three topological relations (Table 1, Supplemental Figure S1).

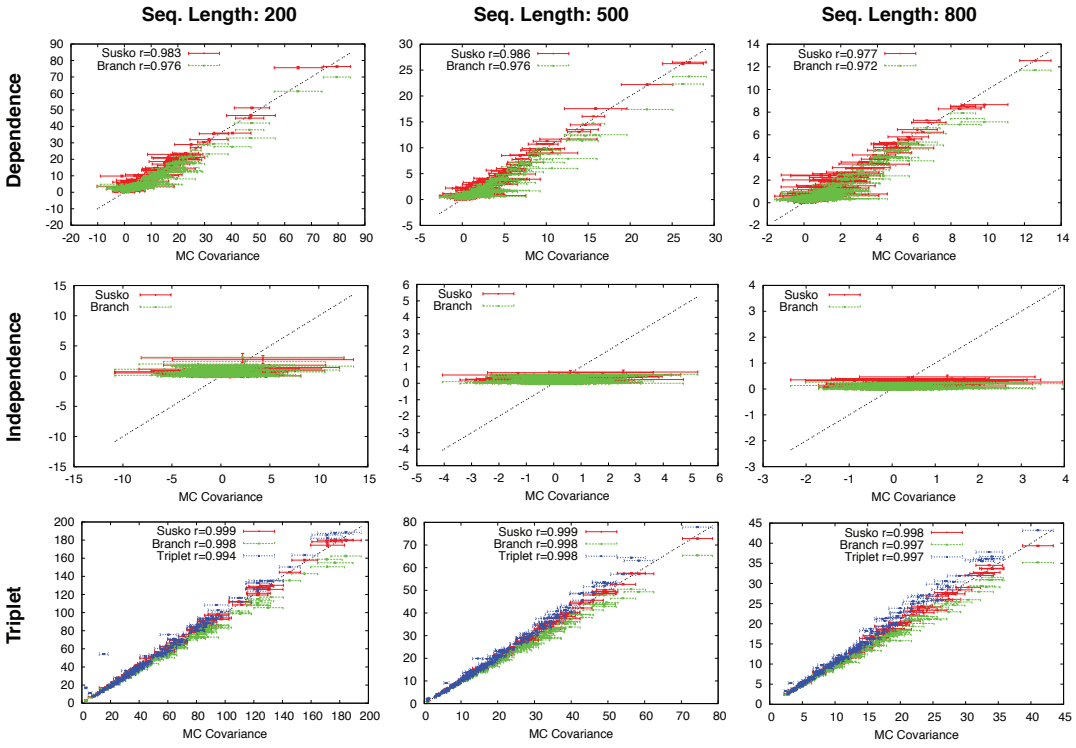

Figure 3. Comparison of Suko-covariance (red), branch-covariance (green) and triple-covariance (blue) with their Monte Carlo counterpart for sequence lengths of {200, 500, 800} amino-acids. Error-bars indicate 95% confidence intervals.

Application to optimal pairwise alignments

We have tested the branch-, anchor-, and triplet-covariance on distances derived from OPAs. The anchor-covariance is and adaptation of the susko-covariance, which was specifically designed to bypass the problem of inconsistent homology inference between OPAs. Since the branch- and triplet-covariance do not directly rely on the sequence data they can also be applied to OPAs.

The branch-covariance has a lower bias than the anchor-covariance for all three topological relations (Figure 4). In the dependence case the differences are minor, though the branch-covariance has a consistently higher correlation with the Monte Carlo covariance. A big difference is visible for the two other topological relations (independence and triplet), where the anchor-covariance's bias is up to twice the branch-covariance's. The superiority of the branch-covariance is also reflected by the average MSEs (Table 2). The triplet-covariance has a greater bias than the branch-covariance for sequences of length 200; for length 500 the two estimators have quantitatively a similar bias, but in opposite directions; and for length 800 the triplet-covariance has clearly a smaller systematic error.

Note that we have tested the anchor- and triplet-covariance under the same simulation conditions in our previous work (Dessimoz and Gil, 2008). They have been reproduced to evaluate the branch-covariance. The results on the triplet- and anchor-covariance reported here are in agreement with our previous results.

	200			500			800		
	D	T	I	D	T	I	D	T	I
Susko	215.96	420.72	125.59	12.74	22.47	8.06	2.76	5.19	1.87
Branch	39.89	255.86	2.37	3.13	16.95	0.12	0.71	4.73	0.03

Table 1. Average mean squared error (MSE) of covariance estimators for dependence (D), triplet (T) and independence (I) case for sequence lengths 200, 500 and 800 amino-acids.

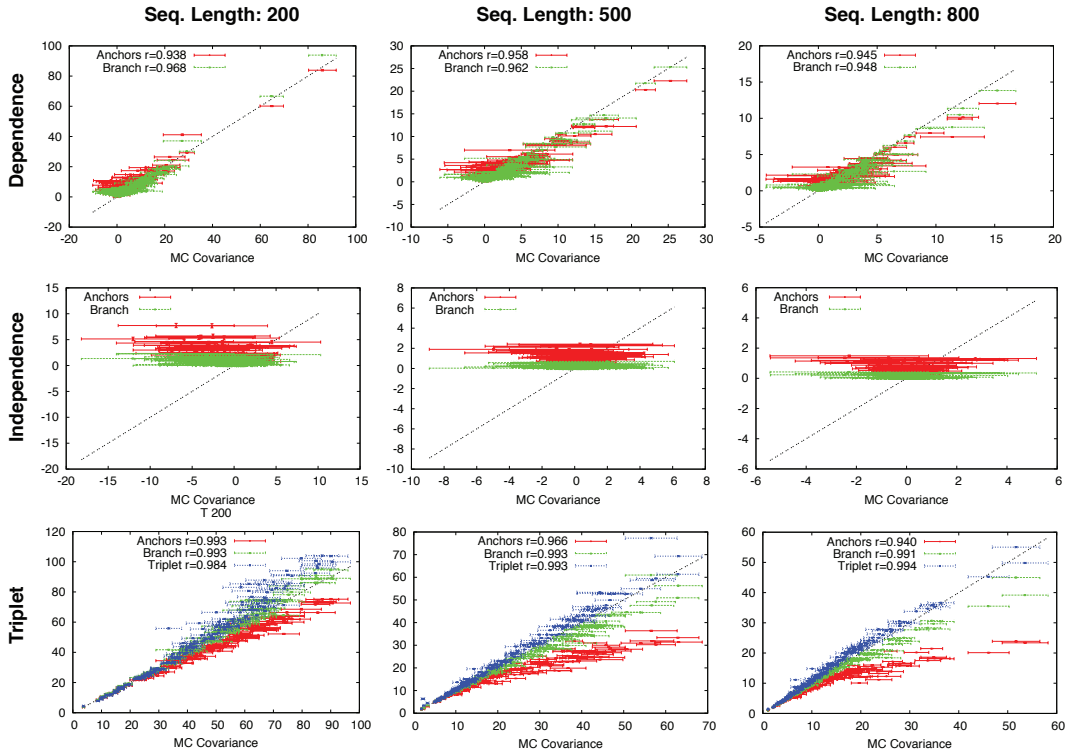

Figure 4. Comparison of anchor-covariance (red), branch-covariance (green) and triplet-covariance (blue) with their Monte Carlo counterpart for sequence lengths of {200, 500, 800} amino-acids. Error-bars indicate 95% confidence intervals.

CONCLUSION

We have presented a fast and general method to estimate the covariance of pairwise ML distances estimates. Our estimator is based on a conjecture (going back to Nei and Jin, 1989) which links the covariance to path lengths on the underlying phylogenetic tree. We have proven it here under a simple symmetric substitution model and formulated it for general models. The estimator is applicable to distances estimated from parametric as well as empirical substitution models and works with both MSAs and OPAs. We have shown by simulation that it has a lower total error than previously published estimators operating directly on the sequence data. Moreover, in contrast to these linear time complexity methods, our estimator runs in constant time in the sequence-length, provided that the pairwise distance information has been precomputed (as in the initial all against all pairwise comparison, customary to most pairwise distance approaches).

We consider the evaluation under the correct model conducted here an important baseline. However, ideal conditions are never met when working with real data, so that as future work the various estimators should be compared in situations where the model-assumptions are violated. Under such conditions, it is conceivable that estimators operating directly on the sequence data outperform the method presented here. The rational being that they extract information from the data at hand –for instance, the actual frequencies

	200			500			800		
	D	T	I	D	T	I	D	T	I
Anchor	159.07	242.56	95.79	11.60	95.62	7.68	3.19	42.87	2.09
Branch	36.19	170.04	3.63	3.99	27.23	0.21	1.38	9.68	0.06

Table 2. Average mean squared error (MSE) of covariance estimators for dependence (D), triplet (T) and independence (I) case for sequences of length 200, 500 and 800 amino-acids in a simulation with indel events.

of aligned character pairs, as opposed to the ones assumed in the substitution model–, and by being more adaptive, they could be more robust to model violations.

REFERENCES

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol*, 229(4):1065–1082.

Bulmer, M. (1991). Estimating the Variability of Substitution Rates. *Genetics*, 123(3):615–619.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. *Evolution*, 21:550–570.

Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., and Gonnet, G. (2005). OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. In McLysath, A. and Huson, D. H., editors, *RECOMB 2005 Workshop on Comparative Genomics*, volume LNBI 3678 of *Lecture Notes in Bioinformatics*, pages 61 – 72. Springer-Verlag.

Dessimoz, C. and Gil, M. (2008). Covariance of maximum likelihood evolutionary distances between sequences aligned pairwise. *BMC Evol. Biol.*, 8(179).

Dessimoz, C., Gil, M., Schneider, A., and Gonnet, G. H. (2006). Fast estimation of the difference between two PAM/JTT evolutionary distances in triplets of homologous sequences. *BMC Bioinformatics*, 7(529).

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, 35:1229–1242.

Fitch, W. and Margoliash, E. (1967). The construction of phylogenetic trees. *Science*, 155:279 – 284.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, 14(7):685.

Gil, M. and Gonnet, G. H. (2009). Phylogenetic tree building methods. In Appel, R. and Feytmans, E., editors, *Bioinformatics - A Swiss Perspective*. World Scientific.

Gil, M., Zanetti, M. S., Zoller, S., and Anisimova, M. (2013). Codonphyml: fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular biology and evolution*, 30(6):1270–1280.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256(5003):1443–1445.

Gonnet, G. H., Hallett, M. T., Korostensky, C., and Bernardin, L. (2000). Darwin v. 2.0: An interpreted computer language for the biosciences. *Bioinformatics*, 16(2):101–103.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro, H., editor, *Mammalian protein metabolism III*, pages 21–132. Academic Press, New York.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518.

Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol*, 2(1):87–90.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–1635.

Mihaescu, R. and Pachter, L. (2008). Combinatorics of least-squares trees. *PNAS*, 105(36):13206–13211.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Nei, M. and Jin, L. (1989). Variances of the Average Numbers of Nucleotide Substitutions Within and Between Populations. *Mol Evol Biol*, 6(3):290–300.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In Gupta S.S, Y. J., editor, *Statistical decision theory and related topics*, pages 1–27. Academic Press; New York.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Roch, S. (2010). Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971):1376–1379.

Stamatakis, A. (2014). Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, page btu033.

Steel, M., Hendy, M., and Penny, D. (1988). Loss of information in genetic distances. *Nature*, 336(6195):118.

Susko, E. (2003). Confidence regions and hypothesis tests for topologies using generalized least squares. *Mol. Biol. Evol.*, 20(2):862 – 868.

Tajima, F. and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Mol Evol Biol*, 1(3):269–285.

Vinh, L. S. and von Haeseler, A. (2004). Iqpnni: moving fast through tree space and stopping in time. *Molecular biology and evolution*, 21(8):1565–1571.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press, London.

SUPPLEMENTAL MATERIAL

S.1 Maple code for derivation in Section *r-state symmetric model*

In this section we give a Maple program which follows the derivation presented in Section *r-state symmetric model*. In particular, the code evaluates the expression on the right-hand side of Equation 16. At the end it compares the derived expression with the conjectured one and demonstrates their equivalence.

```
pm := B*(1-exp(-d/B)): r := 1/(1-B):
Pr_Quartet := proc (i, j, k, l, m)
global r, pm:
    `if`(i = 0, 1-subs(d = ei, pm), subs(d = ei, pm)/(r-1)) *
    `if`(k = 0, 1-subs(d = ek, pm), subs(d = ek, pm)/(r-1)) *
    `if`(m = 0, 1-subs(d = em, pm), subs(d = em, pm)/(r-1)) *
    `if`(j = 0, 1-subs(d = ej, pm), subs(d = ej, pm)/(r-1)) *
    `if`(l = 0, 1-subs(d = el, pm), subs(d = el, pm)/(r-1))/r
end:
f0 := r*(r-1):
f1 := r*(r-1)^2:                          #    \ ei          ej /
f2 := r*(r-1)*(r-2):                      #     \      em      /
f3 := r*(r-1)^2*(r-2):                    #      o------------o
f4 := r*(r-1)*(r-2)^2:                    #     /              \
f5 := r*(r-1)^2*(r-2)^2:                  #    / ek          el \
f6 := r*(r-1)*((r-2)^2+r-1):
f7 := r*(r-1)*((r-2)^2+r-1)*(r-2):
f8 := r*(r-1)*((r-2)^2+r-1)^2:
PAT := [[[0, 0, 0, 0, 1], f0],
       [[0, 1, 0, 1, 0], f1], [[1, 0, 1, 0, 0], f1],
       [[0, 1, 1, 0, 0], f1], [[1, 0, 0, 1, 0], f1],
       [[0, 0, 0, 1, 1], f2], [[0, 0, 1, 0, 1], f2],
       [[0, 1, 0, 0, 1], f2], [[1, 0, 0, 0, 1], f2],
       [[0, 1, 1, 1, 0], f3], [[1, 0, 1, 1, 0], f3],
       [[1, 1, 0, 1, 0], f3], [[1, 1, 1, 0, 0], f3],
       [[0, 1, 0, 1, 1], f4], [[1, 0, 1, 0, 1], f4],
       [[0, 1, 1, 0, 1], f4], [[1, 0, 0, 1, 1], f4],
       [[1, 1, 1, 1, 0], f5], [[0, 0, 1, 1, 1], f6],
       [[1, 1, 0, 0, 1], f6], [[1, 1, 0, 1, 1], f7],
       [[1, 1, 1, 0, 1], f7], [[0, 1, 1, 1, 1], f7],
       [[1, 0, 1, 1, 1], f7], [[1, 1, 1, 1, 1], f8]]:
E_SijSkl := 0:
for p in PAT do E_SijSkl := E_SijSkl+Pr_Quartet(op(p[1]), pm)*p[2] od:
E_SijSkl := simplify(E_SijSkl):
E_Sij := subs(d = ei+em+ej, pm): E_Skl := subs(d = ek+em+el, pm):
cov_IijIkl := (E_SijSkl-E_Sij*E_Skl)*n:
hatd := solve(pm = I_/n, d):
Dhatd_DI := simplify(subs(I_ = n*pm, diff(hatd, I_))):
Dhatd_DI_ij := subs(d = ei+em+ej, Dhatd_DI):
Dhatd_DI_kl := subs(d = ek+em+el, Dhatd_DI):
cov := expand(Dhatd_DI_ij*Dhatd_DI_kl*cov_IijIkl):
cov_conjecture := B*((1-B)*exp(2*em/B)+(2*B-1)*exp(em/B)-B)/n:
proven := evalb(expand(cov_conjecture-cov) = 0); # evaluates to "true"
```

S.2 Maple code for weighted least squares on a quartet

The method of weighted least squares for phylogenetic tree reconstruction from pairwise distance data was first proposed by Cavalli-Sforza and Edwards (1967), and Fitch and Margoliash (1967). The goal is to find an unrooted tree T (topology and branch lengths) that minimizes

$$S(T) = \sum_{i,j} \frac{(t_{ij}(T) - d_{ij})^2}{v_{ij}}, \tag{32}$$

where t_{ij} is the path length on T between leafs i and j, d_{ij} the corresponding input distance and v_{ij} its variance. Sometimes the variances are not known and, therefore, modeled as a function of the distance

estimates. For instance, Fitch and Margoliash assumed that the variances are proportional to the squared distances. When nothing is known about the errors, or if they are assumed to be independently distributed and equal for all observed distances, then all the variances are set to one. This leads to the ordinary least squares method.

The minimization of S is a non-trivial problem. It involves searching the discrete space of unrooted binary tree topologies whose size is exponential in the number of leaves n. The minimization for a given topology is a linear least squares problem formulated by the normal equations. Their algebraic solution involves the computation of a pseudo-inverse. However, elegant combinatorial formulas have been derived for certain variance-structures (Mihaescu and Pachter, 2008).

In Section *Topological relation and branch length* we consider quartet trees. In this case there are only three topological relations so that the enumeration of the tree space is not a problem. We will derive closed form solutions for the direct computation of S and the corresponding optimal branch lengths. The approach implemented in the following Maple script is to take the derivatives of S with respect to the branch lengths $\{e_i\}_{i=1}^{5}$, equate each of the derivatives to zero, solve the resulting system of equations, and substitute the optimized branch lengths $\{\hat{e}_i\}_{i=1}^{5}$ in S:

```
S :=    (e1+e2    -d12)^2/v12        #    \ e1           e3 /
      + (e1+e5+e3-d13)^2/v13        #     \       e5      /
      + (e1+e5+e4-d14)^2/v14        #      o------------o
      + (e2+e5+e3-d23)^2/v23        #     /               \
      + (e2+e5+e4-d24)^2/v24        #    / e2           e4 \
      + (e3+e4    -d34)^2/v34:
dS1 := diff(S,e1): dS2 := diff(S,e2): dS3 := diff(S,e3):
dS4 := diff(S,e4): dS5 := diff(S,e5):
hat_e := solve({dS1=0, dS2=0, dS3=0, dS4=0, dS5=0},
                {e1,e2,e3,e4,e5}):
Sopt := simplify(subs(hat_e,S));
hat_e5 := simplify(rhs(hat_e[5]));
```

The output of the script corresponds to Equations 29 and 30.

S.3 Supplemental Figure

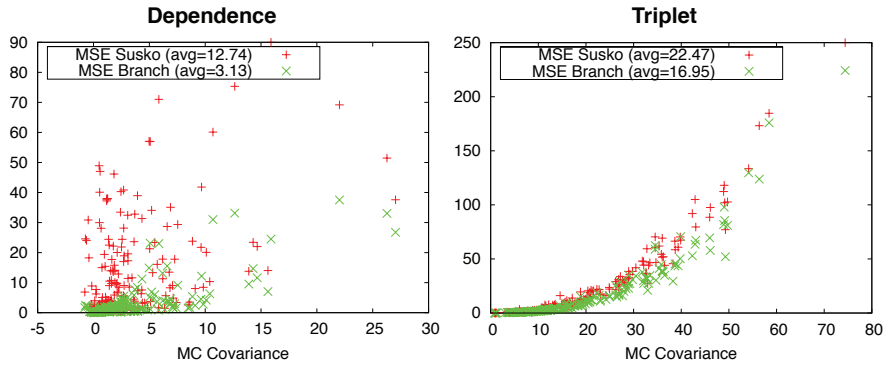

Figure S1. Mean squared error (MSE) of covariance estimators for dependence and triplet case as a function of the Monte Carlo covariance for sequences of length 500 amino-acids. The branch-covariance (green) has a lower MSE than the Susko-covariance estimator (red).