Syst. Biol. 45(2):127–134, 1996

CONFIDENCE INTERVALS FOR THE DIVERGENCE TIME OF TWO CLADES

MIKE A. STEEL,[1] ALAN C. COOPER,[2] AND DAVID PENNY[3]

[1]*Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand; E-mail: m.steel@math.canterbury.ac.nz*
[2]*National Zoological Park, Smithsonian Institution, Washington, D.C. 20008, USA*
[3]*School of Biological Sciences, Massey University, Palmerston North, New Zealand*

Abstract.—We describe a simple method for generating tighter confidence intervals for the date of divergence of two monophyletic groups of taxa. This technique exploits the variation that exists within each of two groups that have evolved separately from a common ancestor. We illustrate the method (plus a technique to test the molecular clock hypothesis) using sequence dissimilarity within the orders of ratites and of tinamous (small birds from South America commonly regarded as the closest relative to ratites). [Aligned sequences; clades; divergence times; molecular clock; phylogeny.]

The estimation of divergence times between taxa is an important problem in phylogenetic analysis (Gojobori et al., 1990; Kishino and Hasegawa, 1990; Marshall, 1990). Here, we address the question of how to exploit the knowledge that two groups are monophyletic to obtain better estimates of sequence divergence between the two groups and thereby to obtain better estimates of their divergence time. The problem is that different pairwise comparisons between taxa, one from each group, cannot be treated as statistically independent measurements because of shared history. Yet it seems equally clear that the collection of all these pairwise comparisons should provide more information about divergence time than can any single comparison.

We describe an approach that allows for the construction of confidence intervals for sequence dissimilarity that are always tighter than those obtained by single pairwise comparisons. This approach is valid under standard statistical assumptions, listed here as assumptions 1 and 2. We illustrate the calculations involved with a simple example (Fig. 1) and also with an application to 12S ribosomal RNA sequences within the avian orders of ratites and of tinamous. This example demonstrates how the addition of sequences to each group leads to narrower confidence intervals (Fig. 2). Our approach produces markedly narrower confidence intervals (than those obtained from single pairwise comparisons) only when there is variation in the sequences within each group. However, the full sequences (not just the derived dissimilarities) must be used to obtain the narrower confidence intervals on sequence dissimilarity.

Under two further assumptions (nos. 3, 4), it is possible to use these confidence intervals for sequence dissimilarities to obtain confidence intervals for the number of substitutions separating the two groups. Assumption 4 is closely related to the molecular clock hypothesis, which asserts that the rate of substitution is constant in different lineages and through time (see Zuckerkandl and Pauling, 1965; Kimura, 1983). To test this assumption, we used a simple test that does not depend on the details of a particular underlying substitution model and applied it to the ratite and tinamou sequences.

The problem we consider resembles but is quite different from the problem considered by Rzhetsky et al. (1995), who also sought to obtain improved estimates of phylogenetic parameters by exploiting prior knowledge that groups are monophyletic. Other authors (e.g., Hasegawa et al., 1987; Gojobori et al., 1990; Kishino and Hasegawa, 1990; Tajima, 1993; Zharkikh, 1994) have also considered (different) questions concerned with better estimating the time of separation between sequences.

(a)

```
Taxon 1   gcttagccctaaatccaaatgcttacctaagcattcgcccg
Taxon 2   gcttagccctaaatctaaatgcttacctaacgattcgcccg
Taxon 3   gcttagccctaaatcctgatacttacctaagtatccgccca
Taxon 4   gcttggccctaaatctagatacttacacaagtatccgccta
Taxon 5   gcttagccctaaatcctggtgcttacctaagtacccgccca
```

(b)

Taking A = {1, 2}; B = {3, 4, 5}, we have (with $N = 41$),

Pattern (π)			Pattern (π)		
12 345	X_π	β_π	12 345	X_π	β_π
αβ ααα	1	3	αα ααα	27	0
αα αβα	4	2	αβ αβα	1	3
αα βαβ	1	4	αα βββ	3	6
αα ααβ	2	2	αα ββα	1	4
αβ γγγ	1	6			

(c)

$$d = \frac{1}{2\times3} \times \frac{(1\times3 + 4\times2 + 1\times4 + \ldots)}{41} = 0.2033$$

$$s^2 = \frac{1}{40}\left[\frac{1}{4\times9} \times \frac{(1\times9 + 4\times4 + 1\times16 + \ldots)}{41} - d^2\right]$$

$$= \frac{1}{40}\left[\frac{1}{36} \times \frac{218}{41} - (0.2033)^2\right] = 0.0027$$

FIGURE 1. Computation described by Theorem 1. (a) A subset of the sites ($N = 41$) for five sequences. (b) For these sites, there are nine distinct patterns. These patterns, their frequencies, and their β values are shown, assuming that taxa 1 and 2 lie on one side of the root and taxa 3, 4, and 5 lie on the other. (c) These patterns are used to compute d and the unbiased estimator s^2 for its variance.

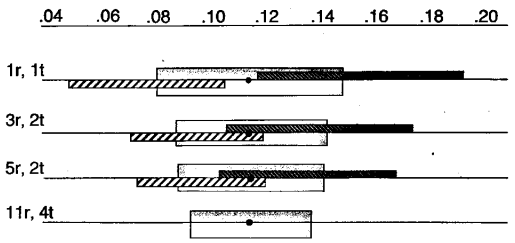

FIGURE 2. The effect of combining taxa to obtain tighter confidence intervals for the (uncorrected) distance between two groups of taxa (e.g., 11 ratites and four tinamous). If a single ratite (r = 1) and tinamou (t = 1) are used, the 2 SD confidence intervals for their sequence dissimilarity are both wide and variable. The average interval (over all choices of one ratite and one tinamou) (▨) is [0.0795, 0.146]; other choices give confidence intervals ranging from [0.047, 0.103] (▨) to [0.115, 0.192] (■). As more taxa are added into the two groups (see line 3r, 2t and line 5r, 2t), the confidence intervals for the average sequence dissimilarity between the groups become narrower and less variable. When all 11 ratites are combined and all four tinamous are combined (11r, 4t), the confidence interval is [0.089, 0.137], whose width is about 70% of the average of the widths of the confidence intervals obtained by taking all single ratite and tinamou comparisons. The reduction in the span of the confidence interval is a major help when testing to see whether a particular divergence occurred, say, before or after the Cretaceous–Tertiary boundary.

The stimulus for this work was testing whether orders of modern birds diverged after the Cretaceous–Tertiary boundary (Feduccia, 1995) or before it. To test the alternatives, calibration points obtained from fossils must be used to estimate rates. Although the major hypotheses could be distinguished, it became clear that some subhypotheses could not be distinguished without a reduction in variance. The reduction formulae reported here should be useful in other studies.

CONFIDENCE INTERVALS

We first obtain a tight confidence interval for the average dissimilarity between taxa from one clade and taxa from the other. We assume that (1) the two clades are identified correctly and (2) sequence sites evolve independently and identically (i.i.d.) according to some stochastic model, M. For assumption 1, we do not need to know the phylogenetic tree connecting the taxa within each clade, just which taxa lie on each side of the root (thus the taxa can be unresolved within each clade). Also, assumption 2 does not compel us to assume that all sites evolve at the same substitution rate; it allows for any distribution of rates across sites, provided the underlying substitution rates are assigned to the sites by an i.i.d. process (Steel et al., 1994; Chang, 1996).

To convert these confidence intervals for average interclade dissimilarities into confidence intervals for the divergence time between the two groups, a confidence interval must be obtained for the number of substitutions, K, separating taxa from one group and taxa from the other. Two further assumptions are required: (3) model M has a correction function ϕ for estimating the expected number of substitutions (K_{ij}) between two sequences i and j from their expected dissimilarity E_{ij}, and (4) K_{ij} takes the same value if i and j are on different sides of the root (a weak form of the molecular clock hypothesis).

Models satisfying assumption 3 include the Jukes–Cantor model and the model described by Tajima and Nei (1982, 1984). For example, the simple Jukes–Cantor model has $K_{ij} = \phi(E_{ij}) = -0.75 \log_e(1 - 4E_{ij}/3)$. Such models are easily modified to handle a variation of rates across sites. For example, with the Jukes–Cantor model, the $\log_e$ function is simply replaced by the functional inverse of the moment generating function of the distribution of rates across sites (this was described explicitly for the gamma distribution by Jin and Nei [1990] and extends directly to more general distributions). Assumption 4 always holds under the molecular clock hypothesis. However, assumption 4 is a slightly more general condition; it would also apply if the substitution rate were constant among the taxa on one side of the root but different from a second substitution rate constant on the other side of the root.

To obtain time scales, it is necessary either to know the substitution rates or to combine the estimated K values with estimated dates for fossils (see Shields and Wilson, 1987; Marshall, 1990; Kornegay et al., 1993). If the substitution rates are

known (or have been estimated to lie in some confidence interval), confidence intervals can be obtained for the dates of divergence between the two groups of taxa from a confidence interval on the K values. For example, suppose the rate matrix of the underlying stationary stochastic matrix is $\mathbf{R}$ and f is the equilibrium frequency of nucleotides (defined by $f\mathbf{R} = 0$ and $\Sigma_{i=1}^{4} f_i = 1$), then the time T from the present to the most recent common ancestor of the two groups is given by

$$T = \frac{-0.5K}{\sum_{i=1}^{4} f_i R_{ii}}$$

(see Rodriguez et al., 1990). For example, under the Jukes–Cantor model with substitution rate λ, $T = K/2\lambda$. Thus, if $[K_1, K_2]$ is a $100(1 - \alpha)\%$ confidence interval for the K value of the two groups, then the interval $[T_1, T_2]$ (where T_i is determined by K_i by the above relationship) is a $100(1 - \alpha)\%$ confidence interval for the time back to the most recent common ancestor of the two groups, T. If case λ is being estimated, perhaps from other data, the corresponding confidence interval for T will be wider. For example, if $[\lambda_1, \lambda_2]$ and $[K_1, K_2]$ are $100(1 - \alpha/2)\%$ confidence intervals for λ and K, respectively, then $[K_1/2\lambda_2, K_2/2\lambda_1]$ is a $100(1 - \alpha)\%$ confidence interval for T. A technique for describing confidence intervals for $[K_1, K_2]$ is given in the corollary to Theorem 1.

Suppose we have two groups of aligned sequences, A and B, and that the true phylogenetic tree has group A on one side of the root and group B on the other. For a sequence i in A and a sequence j in B, let d_{ij} denote the normalized (Hamming) distance between i and j, i.e., the proportion of sites where these two sequences differ. Let d be the average value of d_{ij} over all such choices of i and j:

$$d = \frac{1}{ab} \sum_{i \in A; j \in B} d_{ij},$$

where $a = |A|$, the number of sequences in A, and $b = |B|$.

Even when the sequences evolve independently, the variance of d is not the average of the variances of the d_{ij}'s because of the obvious lack of independence. Nevertheless, one can still obtain (from the sequences but not from the distances alone) a consistent and unbiased expression for the variance of d. Let us call the assignment of states at any single site a *pattern*. Thus, for n aligned DNA sequences there are 4^n possible patterns (although because the sequence length N is generally much smaller than 4^n we usually observe a small subset of these in a given data set). For a collection of aligned sequences, let X_π denote the number of sites in which pattern π occurs, and let β_π denote the number of pairs of sequences (i, j) such that $i \in A$, $j \in B$, and i and j have different states assigned to them by π (for an example, see Fig. 1). Thus, we have

$$d = \frac{1}{ab} \sum_{\pi} \beta_\pi x_\pi,$$

where $x_\pi = X_\pi/N$, the proportion of sites where π occurs. The variance of d can be estimated by the following result.

Theorem 1

Under assumptions 1 and 2, an unbiased and consistent estimator for the variance of d is

$$s^2 := \frac{1}{N-1}\left[\left(\frac{1}{a^2b^2} \sum_{\pi} \beta_\pi^2 x_\pi\right) - d^2\right],$$

where N is the sequence length and the summation is over all patterns π that occur at least once in the sequences. Furthermore, d/s is (approximately) normally distributed (with SD = 1) for $N > 30$.

A calculation illustrating Theorem 1 is given in Figure 1, and a proof of this theorem is given in the Appendix. The statement that s^2 is unbiased as an estimator means that the expected value of s^2 is precisely the variance of d (formally, $E[s^2] = \mathrm{Var}[d]$). As N (the sequence length) becomes large, the ratio $s^2/\mathrm{Var}[d]$ converges to 1 (with probability = 1) so that for sufficiently long sequences s^2 is expected to give an increasingly more accurate estimate of $\mathrm{Var}[d]$ (i.e., s^2 is a consistent esti-

mator). The value of s^2 is less than or equal to $d(1 - d)/(N - 1)$, and this value is attained only when all the sequences in A are equal and all the sequences in B are equal. Thus, variation within the sequences of each group helps to cut down the width of the associated confidence interval for d.

To apply this theorem to obtain a confidence interval for the expected number of substitutions that have occurred between a taxon in one group and a taxon in the other (the K value), we obtain a confidence interval for d (by using the theorem) and then transform the limits of this interval by applying a correction transformation ϕ (ϕ is a monotonic function) as is summarized in the corollary to Theorem 1. The confidence interval for K will not generally be symmetric about $\phi(d)$ because of the nonlinearity of ϕ. For distantly related sequences, the interval may have a very large upper bound and so may only be useful for providing a lower bound on K.

Corollary

Under assumptions 1–3, a $100(1 - \alpha)\%$ confidence interval for the K value of the two groups is $[\phi(d - sz_\alpha), \phi(d + sz_\alpha)]$, where s is given by the theorem and where z_α is the value beyond which the standard normal density has area α.

An Example Using Birds

The theorem was applied to two groups of birds based on 358 third-domain 12S ribosomal RNA sites (the data derived from Cooper et al. [1992] with three additional taxa). The two groups consisted of 11 ratites (two rheas, three kiwis, three moas, one emu, one cassowary, one ostrich) and four tinamous. By applying the theorem on the assumption that the 11 ratites are on one side of the root of the tree and the tinamous are on the other, we obtained a confidence interval for d ($d \pm 2$ SD) of [0.0890, 0.1366] (i.e., 0.1128 ± 0.0238). Applying a Jukes–Cantor correction, the above corollary gives a confidence interval for K of [0.095, 0.593]. However, if we just take individual comparisons of pairs (one ratite and one tinamou) the confidence intervals for d are considerably wider and quite variable (see Fig. 2). As we add more ratites and tinamous into the two groups, there is a steady progression in both the variability and the reduction of the width of the confidence intervals for different selections of taxa. These confidence intervals are for the average dissimilarities (d) between the two groups, not for the corrected K values.

TEST OF A MOLECULAR CLOCK

To justify this approach, it is useful to have a test of the molecular clock hypothesis (or more generally assumption 4) because this is an important assumption in estimating divergence time. A test that does not depend too much on the type of substitution model being considered is desirable. Goldman (1993) described a likelihood-based test of the molecular clock hypothesis assuming that the sequences evolve according to certain stochastic models. Wu and Li (1985) also described a "relative rates" test based on pairwise comparisons, but this test is also dependent on a particular substitution model. Here, we give a simpler test, which is valid for more general models (although it is likely to be less discriminatory than a likelihood test) for three taxa. It can in principle be extended to more than three taxa, but we do not explore this here. Our test is based on the simple observation that for any stationary model (see Rodriguez et al., 1990) a molecular clock (or more generally just assumption 4) implies that the expected value of d_{ij} takes the same value for any pair of taxa i and j that lie on opposite sides of the root of the phylogenetic tree being considered.

Suppose a molecular clock applies to a collection of N sites that evolve i.i.d. Suppose that the rooted tree connecting taxa i, j, and k places i and j on the same side of the root and k on the other side. Under a molecular clock, the expected value of $\delta = d_{ik} - d_{jk}$ is zero. To test this hypothesis, under assumptions 1 and 2, δ is approximately normally distributed because δ is a sum of many i.i.d. random variables (one

for each site), and so the central limit theorem applies. Thus, we need only estimate the variance of δ to perform a significance test using a standard normal table. We can obtain an unbiased estimate of Var[δ] as follows (proof supplied in the Appendix).

Theorem 2

Under assumptions 1 and 2, an unbiased and consistent estimator of the variance of $\delta = d_{ik} - d_{jk}$ is

$$s_1^2 := \frac{(d_{ij} - \delta^2 - d_{ijk})}{N - 1},$$

where d_{ijk} is the proportion of sites where i, j, and k are in three different states. Furthermore, under the molecular clock hypothesis (or more generally, assumption 4), δ/s_1 has (approximately) a standard normal distribution for $N > 30$.

As a simple application, we examined data from two ratites, ostrich and *Megalapteryx* (an extinct moa), which have 27 site differences between them, and from a crested tinamou, which has 41 and 37 site differences from the two ratites, respectively. There is just one site where all three taxa have different states, and because $N = 358$, we have $d_{ijk} = 1/358$, $\delta = 4/358$, and $d_{ij} = 27/358$. Thus, $s_1 = 0.01425$, giving $\delta/s_1 = 0.784$, so this test does not reject the molecular clock hypothesis for these three taxa (the probability of obtaining a deviation of at least 0.784 under a standard normal distribution is not particularly small; standard normal tables give the value as 0.43).

CONCLUSION

We have described a general approach to estimating sequence dissimilarity that exploits knowledge of monophyly to obtain narrower (and therefore more informative) confidence intervals than those obtained by single pairwise comparisons. Applying this approach to bird sequences, the combination of several sequences for two monophyletic groups can lead to a considerable reduction in the width of confidence intervals (Fig. 2). Tighter confidence intervals for sequence dissimilarity lead to narrower confidence intervals on the divergence time between the two groups, providing that a molecular clock holds. We have presented and applied a new and simple test of the molecular clock that does not depend on the fine detail of a particular model for its validity. The formulae reported here should be useful in other studies. The approach presented here could be extended to models that violate assumption 3, i.e., models for which the K value for two sequences depends on more than just their dissimilarity score.

ACKNOWLEDGMENTS

We thank Andrey Zharkikh and the other referee for their useful comments.

REFERENCES

CHANG, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134:189–215.

COOPER, A. C., A. MOURER-CHAUVIRÉ, G. K. CHAMBERS, A. VON HAESELER, A. C. WILSON, AND S. PÄÄBO. 1992. Independent origins of New Zealand moas and kiwis. Proc. Natl. Acad. Sci. USA 89:8741–8744.

FEDUCCIA, A. 1995. Explosive evolution in Tertiary birds and mammals. Science 267:637–638.

GOJOBORI, T., E. N. MORIYAMA, AND M. KIMURA. 1990. Statistical methods for estimating sequence divergence. Methods Enzymol. 183:531–550.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1987. Man's place in Hominoidea as inferred from molecular clocks of DNA. J. Mol. Evol. 26:132–147.

JIN, L., AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82–102.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munrow, ed.). Academic Press, New York.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge, England.

KISHINO, H., AND M. HASEGAWA. 1990. Converting distance to time: Application to human evolution. Methods Enzymol. 183:550–570.

KORNEGAY, J. R., T. D. KOCHER, L. A. WILLIAMS, AND A. C. WILSON. 1993. Pathways of lysozyme evolution inferred from the sequences of cytochrome *b* in birds. J. Mol. Evol. 37:367–378.

MARSHALL, C. R. 1990. The fossil record and estimating divergence times between lineages: Maximum divergence times and the importance of reliable phylogenies. J. Mol. Evol. 30:400–408.

RODRIGUEZ, F., J. L. OLIVER, A. MARIN, AND J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

RZHETSKY, A., S. KUMAR, AND M. NEI. 1995. Four-cluster analysis: A simple method to test phylogenetic hypotheses. Mol. Biol. Evol. 12:163–167.

SHIELDS, G. F., AND A. C. WILSON. 1987. Calibration of mitochondrial DNA evolution in geese. J. Mol. Evol. 24:212–217.

STEEL, M. A., L. A. SZÉKELY, AND M. D. HENDY. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comput. Biol. 1:153–163.

TAJIMA, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 10:677–688.

TAJIMA, F., AND M. NEI. 1982. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. 18:115–120.

TAJIMA, F., AND M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1:269–285.

WILKS, S. 1962. Mathematical statistics. John Wiley and Sons, New York.

WU, C. I., AND W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA 82:1741–1745.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

ZUCKERKANDL, E., AND L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pages 97–166 *in* Evolving genes and proteins (V. Bryson and H. J. Vogel, eds.). Academic Press, New York.

Received 24 May 1995; accepted 1 December 1995
Associate Editor: David Cannatella

APPENDIX
PROOFS OF THEOREMS

Suppose the random vector $[X_i]$ has a multinomial distribution with parameters N and $[f_i]$, and let

$$x_i = \frac{X_i}{N}, \qquad Y = \sum_i \alpha_i x_i.$$

Then

$$S = \frac{1}{N-1}\left(\sum_i \alpha_i^2 x_i - Y^2\right)$$

is an unbiased estimator for Var[Y], i.e.,

$$E[S] = \text{Var}[Y]. \tag{1}$$

Furthermore,

$$S/\text{Var}[Y] \text{ converges to 1 in probability (as } N \to \infty). \tag{2}$$

To derive Equation 1, apply the indentity

$$E[x_i x_j] = f_i f_j\left(1 - \frac{1}{N}\right) + R_{ij},$$

where $R_{ij} = 0$ if $i \neq j$, and $R_{ii} = f_i/N$ (see Wilks, 1962), to obtain

$$\begin{aligned}\text{Var}[Y] &= \sum_{i,j} \alpha_i\alpha_j \text{Cov}[x_i, x_j] \\ &= \frac{1}{N}\left[\sum_i \alpha_i^2 f_i - \left(\sum_i \alpha_i f_i\right)^2\right], \\ E[S] &= \frac{1}{N-1}\left[\sum_i \alpha_i^2 f_i - \sum_{i,j} \alpha_i\alpha_j E[x_i x_j]^2\right] \\ &= \frac{1}{N}\left[\sum_i \alpha_i^2 f_i - \left(\sum_i \alpha_i f_i\right)^2\right],\end{aligned}$$

again by the above identity. This establishes Equation 1. The formula just given for Var[Y] also gives Equation 2 because we have that $[x_i] \to [f_i]$ in probability (as $N \to \infty$). We apply these results to prove the two theorems as follows.

Theorem 1

We have

$$d_{ij} = \sum_\pi \delta_\pi(i, j) x_\pi,$$

where

$$\delta_\pi(i, j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are assigned} \\ & \text{different states by } \pi \\ 0 & \text{otherwise.} \end{cases}$$

Now,

$$\begin{aligned}\sum_{i\in A; j\in B} d_{ij} &= \sum_{i\in A; j\in B} \sum_\pi \delta_\pi(i, j) x_\pi \\ &= \sum_\pi \left[\sum_{i\in A; j\in B} \delta_\pi(i, j)\right] x_\pi = \sum_\pi \beta_\pi x_\pi\end{aligned}$$

Thus,

$$d = \frac{1}{ab}\sum_\pi \beta_\pi x_\pi,$$

and by the i.i.d. assumption, $[X_\pi]$ has a multinomial distribution, so we obtain Theorem 1 by Equation 1 by letting i range over the patterns and setting $\alpha_\pi = \beta_\pi/ab$. Consistency follows from Equation 2, and the central limit theorem (Wilks, 1962) implies that d has an (approximately) normal distribution.

Theorem 2

Partition the sequence sites into five classes, C_1, $C_2, \ldots, C_5$, by considering the patterns on sequences i, j, and k as follows:

C_1: i, j, and k are in the same state;
C_2: i and j are in the same state, k is in a different state;
C_3: i and k are in the same state, j is in a different state;
C_4: j and k are in the same state, i is in a different state;
C_5: i, j, and k are in three different states.

Let $c_i = |C_i|/N$. Thus, $c_5 = d_{ijk}$, and $c_1 + c_2 + c_3 + c_4 + c_5 = 1$. Now,

$$d_{ij} = c_3 + c_4 + c_5$$

$$d_{ik} = c_2 + c_4 + c_5$$

$$d_{jk} = c_2 + c_3 + c_5$$

and because

$$d_{xy} = \sum_{\pi:\pi(x)\neq\pi(y)} x_\pi$$

(for $x, y \in \{i, j, k\}$) we have

$$\delta = \sum_\pi \lambda_\pi x_\pi$$

where

$$\lambda_\pi = \begin{cases} 1 & \text{if } \pi \in C_4 \\ 0 & \text{if } \pi \in C_1 \cup C_2 \cup C_5 \\ -1 & \text{if } \pi \in C_3. \end{cases}$$

Note that

$$\sum_\pi \lambda_\pi^2 x_\pi = c_3 + c_4 = d_{ij} - c_5 = d_{ij} - d_{ijk}.$$

Thus,

$$\frac{1}{N-1}\left(\sum_\pi \lambda_\pi^2 x_\pi - \delta^2\right) = \frac{(d_{ij} - \delta^2 - d_{ijk})}{N-1} = s_1^2.$$

Hence,

$$E[s_1^2] = E\left[\frac{1}{N-1}\left(\sum_\pi \lambda_\pi^2 x_\pi - \delta^2\right)\right]$$

$$= \mathrm{Var}[\delta]$$

by Equation 1, as claimed. Consistency now follows from Equation 2, and the central limit theorem (Wilks, 1962) implies that δ has an (approximately) normal distribution.