LA-UR-00-929

Approved for public release; distribution is unlimited.

Title: Comparison of Signature Pattern Analysis Methods in Molecular Epidemiology

Author(s): Tom Burr, William S. Charlton, and William D. Stanbro

Submitted to: 2000 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences
Las Vegas, Nevada, June 26-29, 2000

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

Comparison of Signature Pattern Analysis Methods in Molecular Epidemiology

Tom Burr
Safeguards Systems Group, NIS-7, Mail Stop E541
Los Alamos National Laboratory
Los Alamos, NM 87545
Ph:(505) 665-7865, fax: (505)667-7626
tburr@lanl.gov

William S. Charlton
Safeguards Systems Group, NIS-7, Mail Stop E541
Los Alamos National Laboratory
Los Alamos, NM 87545
Ph:(505)665-8576, fax: (505)667-7626
wcharlton@lanl.gov

Willam D. Stanbro
Safeguards Systems Group, NIS-7, Mail Stop E541
Los Alamos National Laboratory
Los Alamos, NM 87545
Ph: (505)667-6779, fax(505)667-7626
wstanbro@lanl.gov

Corresponding author: Tom Burr, to present paper if accepted
Keywords: signature pattern analysis, molecular epidemiology

Abstract

We consider the supervised learning problem of assigning test influenza sequences to their correct group (where the group is the host species). Assume that training cases (influenza sequences and their group labels) are available, usually via estimates of phylogenetic (evolutionary) trees as a special case of unsupervised learning. We compare three supervised learning methods: (1) a published signature pattern analysis (VESPA) approach; (2) an unpublished Bayesian approach that assumes sites are independent, and (3) a nearest-neighbor approach with flexible evolutionary distance measures. Although the Bayesian approach has the attractive feature of reporting estimated probabilities for each group for each test sequence, those probabilities are somewhat suspect because of the site-independence assumption that is difficult to remove. We investigate the impact of this independence assumption and show that it can be conservative or anti-conservative (meaning that it leads to either overstating or understating the separability of the groups). The VESPA approach also assumes site independence, but it has the advantage of allowing for dependence among sequence scores in a way that can easily be estimated, as we illustrate. Finally, the distance-based method is always a strong contender, and especially in this case because of the ease of incorporating evolutionary models into the distance measure. All three methods are of potential use on similar problems, with no single method emerging as the clear winner. We compare conclusions under the three approaches for sequences from the Nucleoprotein (NP) gene of the human influenza RNA virus from three host species. This data is available from the influenza database maintained at Los Alamos National Laboratory (http://linker.lanl.gov/flu/search_frame.html).

1. Introduction

There are many [1] settings involving hypothesis testing in the context of evolutionary (phylogenetic) trees. A tree (Fig. 1) describes the branching order and branch lengths. Usually, the branch lengths in the path connecting two taxa are related to the expected genetic distance between them. The tree in Fig. 1a includes 85 nucleoprotein (NP) genes of the influenza virus in three hosts (for clarity, only 85 sequences are shown from our full data set of 129, with 44 avian (A), 57 human (H) and 28 swine (S)). In [2] we focused on situations involving testing whether a prespecified group is monophyletic. A group (clade) is monophyletic if it clusters before any taxa from outside the group. That is, in tracing the taxa back in time to the most recent common ancestor (MRCA), if we do not encounter any taxa from outside the group, then the group is a clade. Here, we will consider the similar but distinct question: given sequence groupings that are known to be correct in training data, find methods to assign new test sequences to the correct group, and estimate the misclassification rate. This is a standard application of supervised learning (also called pattern recognition or classification) in the special context of DNA sequence data. DNA sequence data offers unique challenges because it is discrete and because distances between sequences should be defined using distance measures that relate to the appropriate evolutionary model. We will use distance measures based on sequence mutation models having 1, 2, and 5 parameters, referred to as Jukes-Cantor (JC-1), Kimura (Ki-2), and Hasagewa (HKY-5) models, respectively [3]. If the main concern is to assign test sequences to the correct group, then assuming success with supervised learning, we could avoid the computationally demanding task of rebuilding the entire tree when test sequences are considered. This strategy has application in molecular epidemiology in rapidly assigned test sequences to the correct group, or labeling them as "not like what we've seen before."

Here we assume that training cases (sequences and their group labels) are available, usually via estimates of phylogenetic trees as a special-case of unsupervised learning [3]. We compare three supervised learning methods: (1) a published [4] signature pattern analysis approach; (2) an unpublished Bayesian approach that assumes sites are independent, and (3) a nearest-neighbor approach with flexible evolutionary distance measures. We assume that all training cases are properly labeled with the correct group assignment, and avoid cases that the tree building stage indicate are potentially mislabeled.

Section two provides additional background. Section three defines the three methods, investigates the impact of assuming that sites evolve independently, and compares the methods on a real data (129 NP sequences in H, S, and A hosts as in Fig. 1). Section four indicates directions for future research.

2. Background

The tree in Fig. 1a is a "point estimate" of the true tree obtained by maximum likelihood (ML). For 85 taxa, there are over 10^{62} possible unrooted topologies (branching orders), and Fig. 1a is the one that the data most supports in some sense (based on a consensus of 100 bootstrap trees, where we sample sites with replacement in the bootstrap [5]).

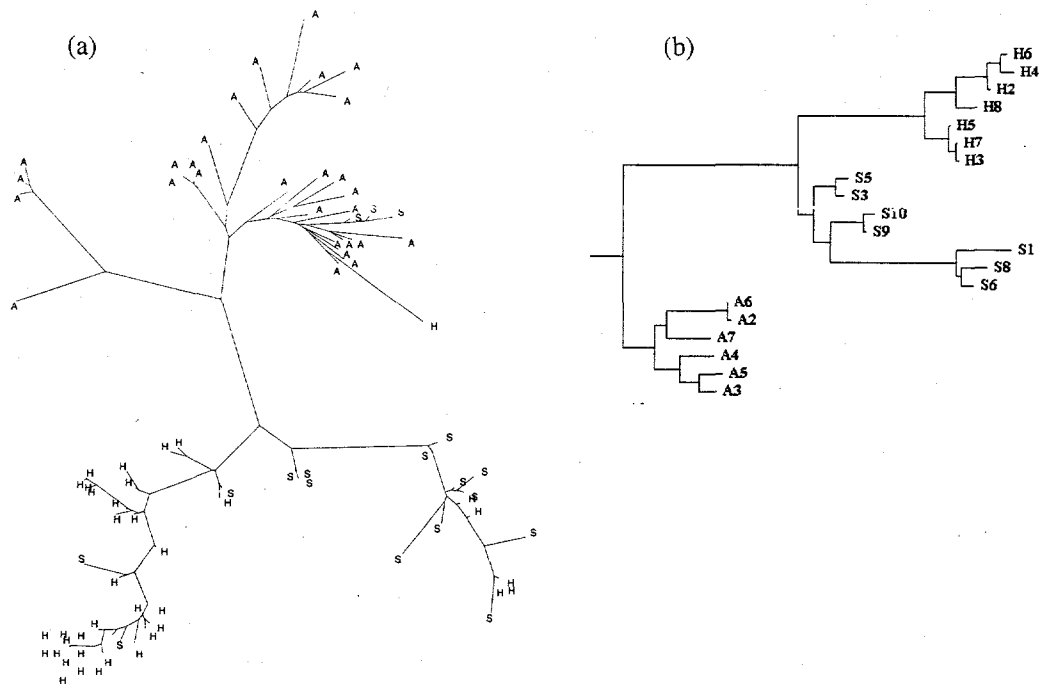

Fig. 1. (a) Consensus tree (of 100 bootstrap samples) using ML for 85 nucleoprotein sequences from human (H), swine (S), and avian (A) hosts , (b) Consensus tree using ML for 18 nucleoprotein sequences from H, S, and A hosts.

Fig. 1a is unrooted (so no time direction is implied) and the internal taxa are unobserved. Fig. 1b is rooted but without using external data, so the root could be placed anywhere. At this stage, we assume the class labels (A, H, and S) are correct, and disregard the crucial tree building (unsupervised learning) stage. However, 14 of these 129 cases are cross-species transmissions. For example, the "A-like" human is one of the poulty-to-human Hong Kong flu cases [2] from 1997.

In [2] we presented a strategy with statistical formalism for assigning a "signature" to a given (monophyletic) group as follows. Given n taxa divided into n_{train} and n_{test} cases, we could either: (A) use all n cases to estimate their phylogenetic tree, or (B) use n_{train} cases to estimate the number of clades present in the n_{train} cases (assumed to be the same as in the n cases). Given c clades defined by the groupings in the n_{train} cases, we could either infer (estimate) the sequence of the MRCA for each clade, or more simply we could define a "consensus" sequence that is the most common nucleotide at each site. Either an inferred MRCA or a consensus sequence could be used as a clade signature for supervised learning. Thus, we have a combination of specialized unsupervised learning (tree building stage) followed by supervised learning (applied to develop "signatures" for each clade). It is much more computationally demanding to invoke (A), so if option (B) provides a reasonable approximation to option (A), then it is likely to be preferred in practice. An application to molecular epidemiology is to use the signatures from (B) for each clade to define the "ordinary" background. Each new test sequence would then be compared to each signature and judged to either belong to the clade having the most similar signature or to be an "outlier." For example, we anticipate that NP sequences maintain 2-5 host-specific groups [6,7] so our 14 cross-species transmissons among the 129 sequences are "outliers" in the sense of being in the "wrong" host.

3. Three Supervised Learning Methods

Once we have selected groups (clades) that are well supported (by the bootstrap for example), we view the problem as a supervised learning problem and define signatures for each group. We have implemented and tested (1) VESPA [4] (viral epidemiology signature pattern analysis); (2) a Bayesian [8] posterior probability method (both of which regard the DNA data as generic categorical data with four categories with no explicit evolutionary model), and (3) k-nearest neighbor [8] using distances defined by the selected evolutionary model.

VESPA To find signature sites for group 1, VESPA searches for sites for which group 1 has high relative frequency (p_{min} = .8 is the default value) of a given nucleotide and the non-group 1 sequences have low relative frequency (p_{max} = .2 is the default value). The same process is repeated to find signatures for all groups. There is no guarantee of finding any

signature sites, but clearly, the more we find with strict p_{min} and p_{max} values, the better the group separation. Test sequences are assigned to the group having the highest p_{VESPA} = percent agreement between the group's signature site values and the test sequence values at the signature sites. The parameters p_{min} and p_{max} can be selected on the basis of performance on training data at separating the groups.

Bayes Our Bayesian approach attempts to assign a probability to each group in typical Bayesian fashion. We have implemented and tested a few variations that differ in whether we pool all the signature sites and all the non-signature sites and in whether we use the concept of signature sites at all, or simply treat each site individually. If we label each site as a signature site or not, we can express the data for each test sequence as the number of agreements with each groups' signature sites. We then have

$$P(H_1 \mid G=1, n_1) = \binom{n_1}{H_1} p^{H_1} (1-p)^{n_1 - H_1} , \qquad (1)$$

where H_1 is the number of agreements ('hits") among the n_1 signature sites for class 1, and p is the average (pooled) relative frequency of the signature values (if site 10 is a signature site for group 1 and the group 1 relative frequencies of A, C, T, and G are .9, .05, .05, and 0, respectively, then A is the signature value for that site). Equation (1) uses more information than VESPA, which uses only the percent agreement with signature sites. For example, if a test sequence agrees with 5 of 10 signature sites for group 1 and with 50 of 100 signature sites for group 2, then Eq. (3) below applied to Eq. (1) gives $P(G = 1|H_1,n_1) >> P(G = 2|H_2,n_2)$. Another alternative is to use all sites "as is," which leads to

$$P(x \mid G=1) = \prod_{i=1}^{n} p_{i,1}(x), \qquad (2)$$

where $p_{i,1}$ is the probability of observing the n-dimensional (n base pairs) vector x if G =1. Either Eq. (1) or (2) can be converted to a posterior probability (via Bayes rule)

$$P(G=1 \mid x) = P(x \mid G=1)P(G=1) / P(x)$$
$$\text{where } P(x) = \sum_i P(x \mid G=i)P(G=i), \qquad (3)$$

where we assume (in both Eq. (1) and Eq. (2)) that the sites are independent to evaluate $P(x|G = i)$. In applying Eq. (3) to Eq. (1), the data x is replaced by a vector $\{(H_1,n_1), (H_2,n_2), \ldots, (H_c,n_c)\}$ containing the number of "hits" between the vector x and each signature pattern, and the number of sites, for each of c clade signatures.

As an important aside, all of the available phylogenetic software also assumes that DNA sites evolve independently, not because it is always a good assumption, but because it is intractable to allow arbitrary types of dependence structures. The impact of the independence assumption during the tree estimation stage is not well known, but [9] demonstrates that the one of the effects of dependence among sites are to "accentuate the long-branch attraction bias" that is well known [3].

Generally, we have had better success with Eq. (2) than with Eq. (1). However, both methods result in probabilities for each group that sum to one, so there is no direct way to detect a test sequence that is not sufficiently similar to any of the recognized groups. VESPA is a good tool for that purpose: we can select a minimum percent with the closest group's signature for the purpose of detecting outliers. Alternatively, we use the "frequentist notion" of rejecting all classes if $P(h \leq H_1)$ for the most probable class is less than some threshold, where h is 0, 1, ..., n_1 and H_1 is the number of signature hits for the most probable class. We plan to address the details of that approach in a separate paper.

k-nearest neighbor Our k-nearest neighbor method uses several distance measures (eg., JC-1, Kim-2, HKY-5 [ ]) to define pairwise distances among sequences. We then use trial-and-error on the training data to choose a good integer k (usually 1 to 10) such that if the distance of a test case to each training case is ranked, and the k nearest neighbors are used to predict the class (using majority rule), there is small misclassification rate on the training data.

3.1 Impact of the Independence Assumption

The independence assumption is made for convenience (and because data limitations often do no permit otherwise) with the hope that conclusions are insensitive to the types of dependence ([9,10,11]) that have been observed among sites. We present two examples that illustrate the concern that conclusions are not robust to all types of dependence.

Example 1: A 2-class problem with p = 2 predictors, with group i having mean μ_i, both classes having the same 2-by-2 covariance matrix Σ, and each group occurring with frequency ("prior probability") 0.5. It is known [12] that the

misclassification probability P for this example is given by $\Phi(-D/2)$, where $D^2 = (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$ and Φ is the cumulative normal distribution. Note that D is the "statistical" distance (which accounts for variance and covariance) between the two group means.

Theorem 1: The difference between the actual (using Σ) P and estimated (using Σ with off-diagonal entries assumed to be 0) misclassification probability P_{est} ranges from –0.5 to 0.5. And the result extends easily to $p > 2$ predictors.

Intrepretation: The independence assumption can be anticonservative ($P_{est} < P$) by 0.5 or conservative ($P_{est} > P$) by 0.5, so conclusions are extremely sensitive to the independence assumption.

Proof: Without loss of generality, assume $\mu_1=\mathbf{0}$, so that μ_2 represents the shift from $\mathbf{0}$. Let EDE^T be the spectral decomposition $\Sigma = EDE^T$, where the columns of E are the eivenvectors and D is diagonal with eigenvalues λ_i. Assuming independence, then Σ is diagonal so $E=I$ (the identity), and $D^2_{indep} = \mu^2_{21}/\lambda_1 + \mu^2_{22}/\lambda_2$. Not assuming independence, then $D^2_1 = (\mu^2_{21}+\mu^2_{22})/\lambda_1$ if $\mu_2 = cE_1$ and $D^2_2 = (\mu^2_{21}+\mu^2_{22})/\lambda_2$ if $\mu_2 = cE_2$. Therefore, $D^2_{indep} - D^2_1$ can be as large (positive or negative) as $|\max\{\mu^2_{21}(1/\lambda_1-1/\lambda_2), \mu^2_{22}(1/\lambda_1-1/\lambda_2)\}|$. The result follows, and we note that whether $P_{est} < P$ or $P_{est} > P$ will depend on how the group mean separation $(\mu_1-\mu_2)$ aligns with the eigenvectors E_1 and E_2.

Example 2: A 2-class problem with $p = 2$ predictors, with predictor 1 (2) being the DNA values (A, C, T, or G) at site i (i + 1). Again the difference between the actual and estimated P can be positive or negative, with $P_{est} - P$ ranging from 0 to 0.5 in the conservative case and $P_{est} - P$ ranging from 0 to at least –0.08 (empirical proof below) in the anticonservative case. The conservative case difference is easily seen to be as large as 0.5 as follows. Let $p_{1,AA}$ denote the probability that group 1 has an A at sites i and i+1, with similar notation for the other probabilities. Suppose that $p_{1,AG} = p_{1,GA} = 0.5$ (i.e., the only observed patterns in group 1 are AG and GA). Suppose that $p_{2,AA} = p_{2,GG} = 0.5$ (i.e., the only observed patterns in group 2 are AA and GG). Then clearly the groups separate perfectly, giving misclassification probability $P = 0$. However, $P_{est} = 0.5$ (same as random guessing) because the marginal distributions are the same for both groups (A occurs at site i with probability 0.5, as does G, and the same for site i + 1, for both groups). The anticonservative case difference requires a more detailed but straightforward calculation. As an example, let $p_{1,AA} = p_{1,CC} = p_{1,TT} = p_{1,GG} = 0.25$ and $p_{1,AA} = .99$, $p_{1,CC} = .009$, $p_{1,TT} = .0009$, and $p_{1,GG} = .0009$. Then $P_{est} - P = -0.088$ or -0.093 depending on whether the independence assumption is also used to evaluate the average misclassification probability over all cases. It is an open question to establish a lower bound for $P_{est} - P$, but clearly the independence assumption is badly violated in this example, and this can lead to either overstating or understating the group separability, as in example 1.

As we will illustrate, example 1 is relevant here because the percent agreements with each group's signature pattern in VESPA could be approximated by a bivariate normal distribution. Example 2 is relevant for our Bayesian approach because of the sensitivity to the independence assumption. However, we believe that the it more likely for the Bayesian approach to be conservative (this is good news, with $P_{est} - P > 0$, so we tend to overstate our misclassification probability) or at least that it is feasible to investigate whether the type of dependence that leads to an anticonservative case is in effect.

3.2. Results

In all cases, we evaluate the methods on test data that is not used to construct the rules. We consider 129 sequences of 1565 base pairs each from the NP gene of the influenza RNA virus from three host species species (H, S, or A). First we removed the 14 cases that were known [2] to be cross-species transmissions, leaving 129 – 14 = 115 cases. Then we randomly selected approximately 70% from each group to be the training cases, with the remaining as test cases.

This data is available from the influenza database maintained at Los Alamos National Laboratory (http://linker.lanl.gov/flu/search_frame.html). In Fig. 2a (a principle coordinate plot which provides a two-dimensional representation of the data that closely preserves pairwise distances [8]) , all clades are distinct, but the H, S clades are "close." Also, there are several "outliers" in each group. For example, the outlying group of 5 A's in (a) are gulls. There are also outlying H and S cases that the literature ([6], [7]) regards as H-like and S-like, respectively. Figure 2b is the same as Fig. 2a, except uses the HKY5 model to compute distances rather than the JC1 model used in Fig. 2a.

In [2] we compared strategy (A) and (B) and in five repeated analyses of 2, 3, 5, 7, 10, and 20 sequences from each host species, the same four H cases were sometimes (in 1 to 2 of the 5 repeated analyses) classified as S (these four H cases form the small H subgroup in both Figs. 2a and 2b that is near the S group). Otherwise, the remaining cases were all correctly classified. For NP, we concluded that with approximately seven or more taxa from each host, strategy (B) provides a good approximation to strategy (A). As mentioned, for our distance-based method and for tree building, we used several evolutionary models (JC1, Ki2, and HKY5). Also, we applied the models both to all sites and separately to each of the "hot," (antigenic in case of HA) "cold," and "warm" sites (having high, medium, and low mutation rates respectively). Here, we assume we know the correct groupings without building trees so we focus on the supervised learning stage.

All methods could be tuned to give 0 misclassifications among the 79 (30 A, 35 H, 14 S) training sequences, and after doing so, all methods correctly found the 14 cross-species transmissions, plus correctly labeled each of the other 36 test

cases. Details about each method are provided below. And, all methods provide at least qualitative indication of how separable the classes are. Qualitatively, if the percentage of training classes that are "near" a given training case are nearly all the same class for most training cases then the classes separate well in training data and are expected to separated well in testing data.

VESPA. There were no signature sites for S with the default p_{min} and p_{max} values (0.8 and 0.2), but 60 sites for S with $p_{min} = 0.7$ and $p_{max} = 0.3$. Consider the maximum p_{VESPA} over the 3 groups for cases within A, H, and S groups:

A group: 0.6 <= maximum p_{VESPA} <= 0.96, with average of maximum $p_{VESPA} = 0.84$ (121 signature sites)

H group: 0.6 <= maximum p_{VESPA} <= 1.00, with average of maximum $p_{VESPA} = 0.94$ (79 signature sites)

S group: 0.6 <= maximum p_{VESPA} <= 1.00, with average of maximum $p_{VESPA} = 0.94$ (60 signature sites)

The correlation matrices (required for complete assessment as illustrated in Ex. 1) had large negative entries ranging from –0.46 to -.88, and one large positive entry (.80). Regardless of whether we assume multivariate normality of the 3-dimensional signature scores, these correlation matrices are essential to fully characterize the groups' separability. Fig. 2 indicates that the raw percentages have approximate normality within each group, except for the possibility of minor "substrains" that suggest some groups to be mixtures. Here if we assume normality within each group (so that Example 1 applies), then under the independence assumption, for example, the H and S groups are expected to be completely separated in that the expected misclassification probability is $\Phi(-7.2)$, which is essentially 0. If we instead use the estimated covariance matrix, then the expected misclassification probablility is $\Phi(-6.5)$, which is also essentially 0. Therefore, the (incorrect) independence assumption gave an anticonservative result, but was essentially the same as the correct result. Note that a strong advantage of VESPA is that there is no need to assume independence among signature scores because we can estimate the covariance among signature scores using training data. Also note that the VESPA analysis actually becomes a Bayesian analysis when we apply it as just illustrated (if we assume normality, then the supervised learning method illustrated in Example 1 is a Bayesian method [12]).

Bayes. To avoid problems with obtaining relative frequencies of 0 or 1, we added ½ count to observed absolute frequencies and 1 count to the number of taxa. We found that the method that used average percentage agreements over all signature sites was not effective, resulting in 4 – 7 misclassified cases. The method that used each signature sites actual percentage agreement results in 0 misclassified cases.

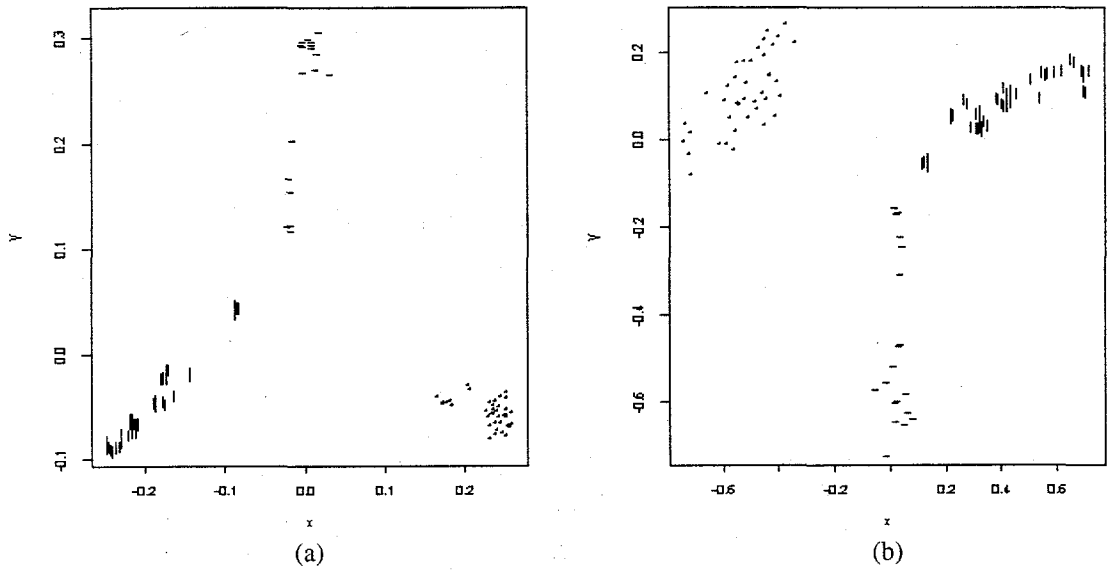

Fig. 2. 115 NP cases: "*" = A, "-" = S, "|" = H. (a) Principal coordinate plot of distances from JC1. (b) Principal coordinate plot distances from HKY5. All clades are distinct, but the H, S clades are "close."

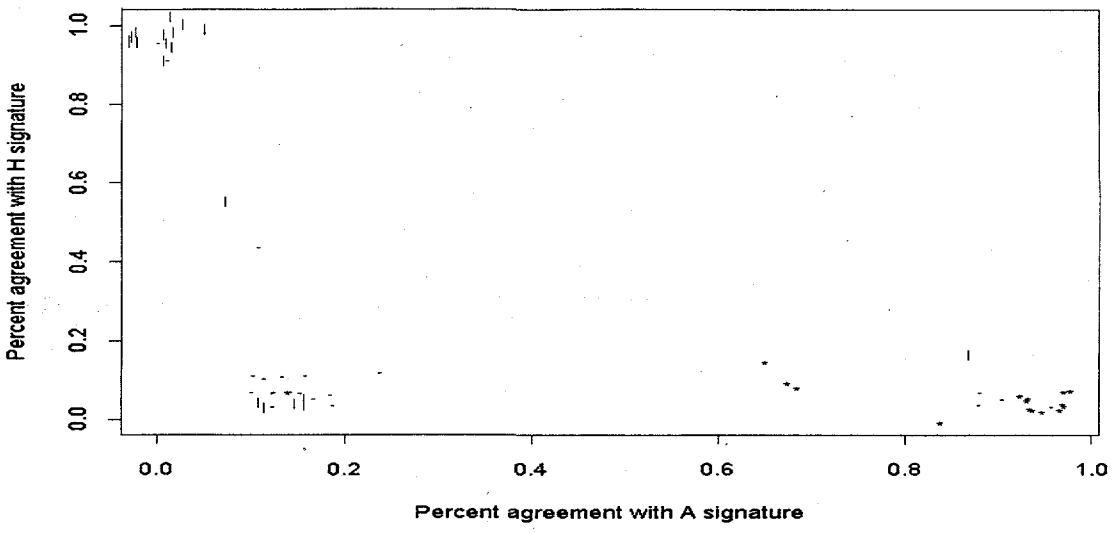

Fig. 3. Percent agreement with H signature vs percent agreement with A signature for H (|), A (*), and S(-). For clarity, only a subset of the 129 cases is shown (15 of each of A, H, S, plus 14 "cross-species" transmissions).

The confidence using VESPA can be assessed by how close the observed maximum percent agreements typically are to their expected values. With the Bayes methods (Eqs. (1) and (2), which both appeal to Eq. (3)), the highest class probabilities were nearly 1. The VESPA maximum percent agreement was quite low (giving a frequentist p-value of .0001 or less because each clade had at least 60 signature sites for p_{min} = .7 and p_{max} =.3) for 20 of the 115 "correct host" sequences. These same 20 sequences are all "on the edges" of the clades in Figs. 2a and 2b. Because the Bayesian probabilities sum to one, the maximum probability for each of these 20 cases was nearly 1. Although we know from the origins of the data that all 20 of these "fringe" cases are "in the correct host," the VESPA maximum percentages indicate that they are somehow different, which is an advantage over our current Bayesian approach. For example, five of the cases are the five edge cases in the A clade that are all gulls from one study. In at least one example, the method of cell culture (growth in chicken eggs) was different, which could have caused some base pair changes. Finally, we know from Example 2 that the independence assumption in Eq. 1 and Eq. 2 can be conservative or anticonservative, which is a disadvantage of this method. However, we expect that in most cases the independence assumption will be conservative. And, in this case a preliminary correlation analysis (using 0-1 valued data by converting A and G to 0 (purines) and C and T to 1 (pyrimidines)) did not reveal any site-to-site dependence at any lag from 1 to 10.

Nearest neighbor. Provided k = 10 (approximately), there were no misclassifications on either the training or testing data. The percentage of training cases among the k = 10 nearest neighbors (using any of our three distance measures) that were the majority class ranged from 0.6 to 1.0 for all three classes, with the mean for A and H approximately 0.95 and for S approximately 0.72. These percentages can be interpreted as the probability of getting the correct classification, so having 0 misclassified out of 129 (all training and testing cases) was unexpectedly good, especially for S. For k = 1. 2, and 5, there were approximately 15, 13, and 7, respectively, misclassified cases (difference results with different distance measures).

4. Conclusions/Future Directions

We believe each of the three approaches has merit. Because of the high Bayesian probabilities (and despite the occasional low VESPA maximum percent agreements) there is strong indication that the clade assignments are correct, although a challenge in this type of application is to ensure that the training cases have the correct group assignments. In this case, the 14 cross-species transmissions are all of the "wrong host" variety, so that for training purposes it would be better to label the "avian-like" human as avian (and similarly for the year labels in HA). Because of the within-group scatter, there are some "fringe" cases that could be regarded as "separate groups" but our current level of assessment ignores subgroups within each main group.

For NP, all 14 of the cross-species transmissions were identified as being in the "wrong host" via the chosen versions of any of the three methods, so the finer points regarding sensitivity to the independence assumption do not cause confusion. However, we recognize the potential sensitivity to the independence assumption and have demonstrated (via Examples 1 and 2) that a complete analysis must consider dependence among scores. It is far simpler to do that with VESPA (or implicitly with k-nn) than with the Bayes methods in Eq. (1) or (2). But if we use VESPA scores as input to a Bayesian method, then it is feasible to estimate the covariance matrix and thereby defend probability estimates. We believe (work in

progress) that the type of dependence that is likely to be in effect for our Bayes approach under Eq. (1) or (2) will lead to a conservative assessment of group separability under the (false) independence assumption, although Examples 1 and 2 indicate that in general the independence assumption can be conservative or anticonservative.

References

[1] Huelsenbeck, J. and Rannala, B., "Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context," *Science*, **276**, 227-232 (1997).

[2] Burr, T., Skourikhine, A., Bruno, W., and Macken, C. "Confidence Measures for Evolutionary Trees: Applications to Molecular Epidemiology," LA-UR-99-4717, *IEEE Proceedings of the International Conference on Information, Intelligence and Systems*, Washington, D.C., November 1-3, 1999.

[3] Swofford et. al., "Phylogeny Inference," Molecular Systematics, second edition, eds Hillis et. al., (1996).

[4] Korber, B. and Myers, G., "Signature Pattern Analysis: a Method for Assessing Viral Sequence Relatedness," AIDS Research and Human Retroviruses **8**, 1549-1560 (1992).

[5] Efron, B., Halloran, E. and Holmes, S., "Bootstrap Confidence Levels for Phylogenetic Trees," *Proc. Natl. Acad. Sci. USA* **93**, 13429 (1996).

[6] Gammelin, M, Mandler, J., and Schholtissek, C. "Two Subtypes of Nucleoproteins (NP) of the Influenza Viruses," Virology **170**, 71-80 (1989).

[7] Gorman, O. et. al., "Evolution of Influenza A Virus NP Genes: Implications for the Origins of H1N1 Human and Classical Swine Viruses," Journal of Virology **65** (7), 3704-3714 (1991).

[8] Venables, W., and Ripley, B., Modern Applied Statistics with S-PLUS, 2nd ed, Springer-Verlag: NY, (1997).

[9] Feng, Y. "The effects of dependence among sites in phylogeny reconstruction,'' Master's Thesis, Mathematics Dept., University of Southern California, (1995).

[10] Borodovsky, M, Sprizhitsky, Y., Golovanov, E., and Alexandrov, A., "Statistical patterns in the primary structures of functional regions in the genomse of E. coli: nonuniform Markov models," Mol. Biol., 20, 1024-1033 (1986).

[11] Watterson, G., "A stochastic analysis of three viral sequences," Mol. Biol. Evol., 9, 666-677 (1992).

[12] Johnson, R., and Wichern, D., Applied Multivariate Statistical Analysis, 2nd ed, Prentice Hall: Englewood Cliffs, NJ, (1988).