

Revised 5/31/00

Assessing Dissimilarity of Genes by Comparing Their RNase A Mismatch Cleavage Patterns

Andrey Rzhetsky,^{*} Joaquín Dopazo,[†] Eric Snyder,[‡] Charles A. Dangler,^{‡§} and Zé Ayala^{*}

^{*}Institute of Molecular Evolutionary Genetics and Department of Biology,

The Pennsylvania State University, University Park, PA 16802, USA

[†]R&D Department, TDI, c/ Condes de Torrealaz 5, 28028, Madrid, SPAIN

[‡]Department of Veterinary Science, The Pennsylvania State University, University Park,
PA 16802, USA

[§]Massachusetts Institute of Technology, Division of Comparative Medicine, Boston, MA
02139, USA

Running title: Genetic distance from RAMCM patterns

Keywords: RNase A mismatch cleavage method, distance estimation, computer simulation, pseudorabies virus.

Corresponding author:

Andrey Rzhetsky

328 Mueller laboratory

University Park, PA 16802

USA

e-mail: aur1@email.psu.edu

Tel: (814) 865-1034

Fax: (814) 863-7336

ABSTRACT

We propose a simple algorithm for estimating the number of nucleotide differences between a pair of RNA or DNA sequences through comparison of their RNase A mismatch cleavage patterns. In the RNase A mismatch cleavage technique two or more sample sequences are hybridized to the same RNA probe, the hybrids are partially digested with RNase A, and the digestion products are compared on the same electrophoretic gel. Here we provide an algorithm for converting the numbers of unique and matching bands between the electrophoretic lanes into an estimate of the number of differences between the sequences. Computer simulation indicates that the proposed method yields a robust estimate of the genetic distance despite stochastic errors and occasional violation of certain assumptions. Our study suggests that the method performs best when the distance between the sequences is less than 15 differences. When the sequences under analysis are likely to have larger distances, we advise to substitute one long riboprobe with a set of shorter non-overlapping probes. The new algorithm is applied to infer the proximity of several strains of pseudorabies virus.

INTRODUCTION

The RNase A mismatch cleavage method (RAMCM) is a powerful technique for detecting single base substitutions (see LÓPEZ-GALÍNDEZ *et al.* 1995). This method is based on the observation that single base mismatches present in RNA:RNA or RNA:DNA hybrids can be cleaved with bovine pancreatic ribonuclease (RNase A, see WINTER, YAMAMOTO *et al.* 1985; MYERS, LARIN *et al.* 1985). In RAMCM a radioactively labeled RNA probe is hybridized to either an RNA or DNA sample, and the hybrid molecule is subjected to RNase A digestion. Each sample treated in this way generates a characteris-

tic pattern of electrophoretic bands which is highly reproducible under specified digestion conditions (PERUCHO, 1989), and the similarity between any pair of sequences can be quickly assessed.

RAMCM is particularly useful in screening a large number of similar sequences. At the beginning, this technique was used for detecting single base substitutions in human genes that might have caused tumorigenesis (FORRESTER, ALMOGUERA *et al.* 1987); it was then applied for other purposes such as studying genetic variability in RNA viruses. In this way, the genetic variabilities of influenza virus (LÓPEZ-GALÍNDEZ, LÓPEZ *et al.* 1988), cucumber mosaic virus (OWEN and PAULUKAITIS, 1988), respiratory syncytial virus (GARCÍA, MARTÍN *et al.* 1994), herpes simplex virus (ROJAS, DOPAZO *et al.* 1995), and human immunodeficiency virus (LÓPEZ-GALÍNDEZ, ROJAS *et al.* 1991), among others, have been analyzed. Recent studies have shown that RAMCM generally produces the same phylogenetic classification as other techniques such as restriction fragment length polymorphism (RFLP, see ROJAS, DOPAZO *et al.* 1995), direct sequencing (GARCÍA, MARTÍN *et al.* 1994) and other molecular screening methods (WHO NETWORK FOR HIV ISOLATION AND CHARACTERIZATION, 1994; SÁNCHEZ-PALOMINO *et al.* 1995). Although genetic variability can be analyzed in a greater detail by direct comparison of nucleotide sequences (NEI and JIN, 1989; NEI, 1987), direct sequencing is rather expensive in terms of time and reagents required. Other approaches, such as RLFP, permit large-scale analyses (NEI and LI, 1979), but lack the high resolution provided by nucleotide sequencing. RAMCM combines the advantages of both direct sequencing and RLFP, because many samples can be quickly and inexpensively analysed, and the method is highly sensitive in detecting single base variations in nucleotide sequences.

Despite the apparent utility of the method, virtually no attempts to ascribe genetic

distances between genes to the observed differences between RAMCM patterns were made until recently (DOPAZO, SOBRINO *et al.* 1993). Using a computer simulation, DOPAZO, SOBRINO *et al.* (1993) showed that the number of different fragments between two RAMCM patterns is proportional to the number of differences between the analyzed nucleotide sequences. The relationship holds for a wide range of realistic conditions used in application of RAMCM. Unfortunately, the factor of proportionality appears to be different in each particular case. Thus, in the absence of independent data allowing for calibration of the measurements, only a rough estimate of the genetic distances can be obtained by that method.

Here we present a new and more accurate algorithm for estimating the number of different nucleotides between two sequences with RAMCM. We use computer simulations to test the validity of the method and to define the optimal conditions to apply this technique in large-scale genetics studies.

MODEL

Definitions and parameters. The ultimate goal of our modeling is to relate the observed numbers of electrophoretic bands in RAMCM to the actual number of nucleotide differences between a pair of sequences. That is, we consider two sample sequences, S_1 and S_2 , that are each hybridized to a reference sequence, S_r , to yield hybrids $S_r::S_1$ and $S_r::S_2$ (see Figure 1 B). Comparing two electrophoretic lanes corresponding to the fragments of the digested hybrids (see Figure 1 B, C, and D), we count separately the bands that are common for two lanes (B_{12}) and the bands that are unique for $S_r::S_1$ and $S_r::S_2$ (B_1 and B_2 , respectively, see Figure 1 C and D). Our goal is to estimate the expected number of differ-

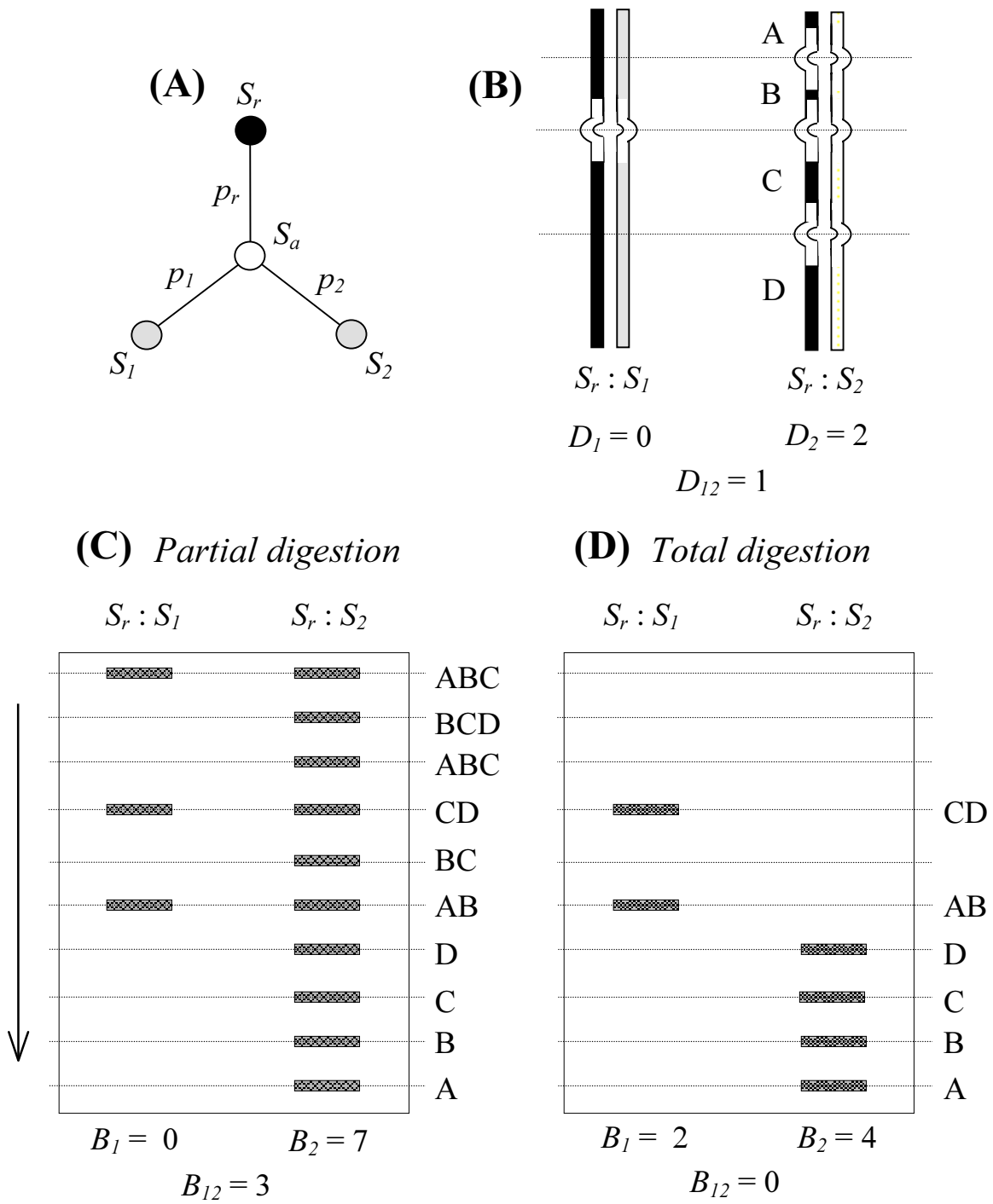


Figure 1.

ences, N_{12} , between sequences S_1 and S_2 from the observed values B_1 , B_2 , and B_{12} .

To derive analytical expressions, we need to make assumptions about the model of nucleotide substitution in sequences S_1 , S_2 , and S_r . Since for hybridization these sequences must be similar with each other and have nearly identical nucleotide frequencies, we can assume that nucleotide substitution is governed by a *stationary* time-reversible Markov model (see KEILSON 1979 for details). We shall use further only two properties common for all stationary time-reversible models: (1) that the expected nucleotide frequencies do not change with time, and (2) that the mathematical description of the substitution process does not change when substitution events are considered “backwards in time” (from the present to the past), which makes the position of the root in a phylogenetic tree immaterial. In other words, labeling nucleotides A, T (U), C, and G by integers 1, 2, 3, and 4, respectively, we can introduce parameters π_1 , π_2 , π_3 , and π_4 , that specify the *expected* nucleotide frequencies *in each* of the sequences S_1 , S_2 , and S_r and in their common ancestors. Further, we can represent the history of origin of these sequences with an unrooted three-sequence tree (see Figure 1 A), denoting by S_a the ancestral sequence corresponding to the interior node of this tree.

We shall also need two additional assumptions regarding the evolution of the sequences: that the only source of evolutionary change in the sequences is nucleotide substitution (*i.e.*, no deletions, insertions, etc.), and that all nucleotide sites of the three sequences accumulate nucleotide substitutions at roughly the same rate.

The actual procedure of digestion of heteroduplexes with RNase A has two important features which also have to be incorporated into the model: (*i*) some mismatches are *undetectable* with RNase A digestion and (*ii*) the digestion can be *partial* or *total* depend-

ing on the conditions of the experiment. The *first property* is evident from the experiments showing that the efficiency of cleavage of a mismatch with RNase A depends both on the actual kind of mispaired bases and the nucleotide sequences flanking this mismatch (PERUCHO 1989; LÓPEZ-GALÍNDEZ, LÓPEZ *et al.* 1988): when a mismatch is present both in $S_p:S_1$ and $S_p:S_2$ heterohybrid molecules at the same site, RNase A either cleaves it *in both* hybrid molecules, or leaves them intact in both. It is not difficult to verify that under a time-reversible stationary model the relative frequencies of different mismatches do not change with time, and the probability of observing a particular sequence flanking a specified mismatch is time-independent. Therefore, we can introduce an additional parameter, P_{CL} , to specify *the probability of cleaving an existing mismatch*, so that on average $(1 - P_{CL}) \times 100$ per cent of the actually existing mismatches would remain *undetected* under our model. The experimentally measured value of P_{CL} is close to 0.45 (PERUCHO 1989). The *second property* applies to the cleavage of *detectable* mismatches. Under *partial* digestion conditions individual heterohybrid molecules with several *detectable* mismatches may not be cleaved at all, may be cleaved only at one of these mismatches, at any two of them, and so on. Since the number of hybrid molecules in the typical sample is very large, the partial digestion generates the set of all possible fragments of the heterohybrid molecule. In the case of the *total* digestion, the same subset of *detectable* mismatches is cleaved in *every* hybrid molecule in a sample. We assume for the moment that all non-homologous heterohybrid fragments can be distinguished by electrophoresis.

Before proceeding to derivation of distance estimators, we need to introduce a number of additional notations which will enable us to shorten the following derivation. First, let S_{ri} , S_{li} , S_{2i} , and S_{ai} stand for nucleotides occupying the i th homologous site in

sequences S_r , S_1 , S_2 , and S_a , respectively. Second, let p_r , p_1 and p_2 define the probabilities of observing $S_{ai} \neq S_{ri}$, $S_{ai} \neq S_{1i}$, and $S_{ai} \neq S_{2i}$, respectively (see Figure 1 B). Third, we need to define the following configurations of sites resulting from pairwise comparisons of sequences S_1 and S_2 with sequence S_r .

$$C_1 = \{S_{1i} \neq S_{ri} \wedge S_{2i} = S_{ri}\}, C_2 = \{S_{1i} = S_{ri} \wedge S_{2i} \neq S_{ri}\},$$

$$\text{and } C_{12} = \{S_{1i} \neq S_{ri} \wedge S_{2i} \neq S_{ri}\}. \quad (1)$$

(Here the symbol ‘ \wedge ’ stands for “and.”) Fourth, we denote by D_1 , D_2 , and D_{12} the *observed* numbers of sites with *detectable* mismatches *and* configurations C_1 , C_2 , and C_{12} , respectively. Finally, by analogy with Equation 1, we introduce eight possible site configurations (denoted here by K_1 , K_2 , ..., and K_8) identifiable by pairwise comparisons of the i th sites of sequences S_r , S_1 , and S_2 with the i th site of the sequence S_a .

$$K_1 = \{S_{ri} = S_{1i} = S_{2i} = S_{ai}\}, K_2 = \{(S_{ri} \neq S_{ai}) \wedge (S_{1i} \neq S_{ai}) \wedge (S_{2i} \neq S_{ai})\},$$

$$K_3 = \{(S_{ri} \neq S_{ai}) \wedge (S_{1i} = S_{2i} = S_{ai})\}, K_4 = \{(S_{1i} \neq S_{ai}) \wedge (S_{ri} = S_{2i} = S_{ai})\},$$

$$K_5 = \{(S_{2i} \neq S_{ai}) \wedge (S_{ri} = S_{1i} = S_{ai})\}, K_6 = \{(S_{ri} \neq S_{ai}) \wedge (S_{1i} \neq S_{ai}) \wedge (S_{2i} = S_{ai})\},$$

$$K_7 = \{(S_{ri} \neq S_{ai}) \wedge (S_{1i} = S_{ai}) \wedge (S_{2i} \neq S_{ai})\}, \text{ and}$$

$$K_8 = \{(S_{ri} = S_{ai}) \wedge (S_{1i} \neq S_{ai}) \wedge (S_{2i} \neq S_{ai})\}. \quad (2)$$

ESTIMATING THE DISTANCE

The forthcoming derivation consists of the following two steps. At the beginning we shall clarify the relationship of the numbers of observed electrophoretic bands, B_1 , B_2 ,

and B_{12} , with the numbers of *detectable* differences, D_1 , D_2 , and D_{12} , between the hybridized sequences (see Figure 1 B). After that, we shall relate the expected values of D_1 , D_2 , and D_{12} with the parameters of the evolutionary tree for sequences S_1 , S_2 , and S_r shown in Figure 1 A, and connect the expected values of B_1 , B_2 , and B_{12} with the expected number of differences, N_{12} , between sequences S_1 and S_2 .

Converting B_1 , B_2 , and B_{12} into D_1 , D_2 , and D_{12} : Partial Digestion. Assuming that all fragments generated by the partial digestion of $S_r:S_1$ and $S_r:S_2$ are identifiable, one can obtain the following simple relationships between the observed numbers of bands and the numbers of *detectable* mismatches in hybrid molecules.

$$B_{12} = 2 D_{12} + \binom{D_{12}}{2} + 1, \text{ and } B_i = 2 D_i + \binom{D_i}{2} + D_i D_{12}, \quad i = 1, 2. \quad (3)$$

Therefore, we obtain the following expressions for computing D_1 , D_2 , and D_{12} from the observed values of B_1 , B_2 , and B_{12} .

$$D_{12} = (\sqrt{8B_{12} + 1} - 3) / 2, \quad \text{and } D_i = (\sqrt{8B_{12} + 8B_i + 1} - \sqrt{8B_{12} + 1}) / 2, \quad i = 1, 2. \quad (4)$$

Total digestion. If all fragments of the total digestion can be identified, the relationship between the numbers of detectable differences and the total numbers of electrophoretic bands in two lanes is as follows.

$$(B_i + B_{12}) = (D_i + D_{12} + 1), \quad i = 1, 2. \quad (5)$$

Although the quantities $(B_1 + B_{12})$ and $(B_2 + B_{12})$ are constant for all fixed values of D_1 ,

D_2 , and D_{12} , the actual values of B_1 , B_2 , and B_{12} can vary depending on the arrangement of mismatches in the hybrid molecules. If the distribution of mismatches is completely random, the probability of observing exactly $B_{12} = k$ shared bands between two electrophoretic lanes is given by

$$\text{Prob}\{B_{12} = k\} = \binom{D_{12} + 1}{k} \binom{D_1 + D_2 - 1}{D_{12} - k} \binom{D_1 + D_2 + D_{12}}{D_{12}}^{-1} \quad (6)$$

It is then easy to find that the expected value of B_{12} is as follows.

$$E(B_{12}) = D_{12} (D_{12} + 1) / (D_1 + D_2 + D_{12}). \quad (7)$$

Finally, combining the last expression with Equation 5 yields

$$D_i = [3 E(B_{12}) + 2 E(B_i) - 1 - \sqrt{9 E(B_{12})^2 + 4 E(B_{12}) \{E(B_1) + E(B_2)\} - 6 E(B_{12}) + 1}] / 2,$$

$$D_{12} = [-E(B_{12}) - 1 + \sqrt{9 E(B_{12})^2 + 4 E(B_{12}) \{E(B_1) + E(B_2)\} - 6 E(B_{12}) + 1}] / 2, \quad (8)$$

where $i = 1, 2$. Therefore, we can obtain estimates of D_1 , D_2 , and D_{12} by replacing the expected values of B_1 , B_2 , and B_{12} in Equation 8 with their observed values. Another way to estimate these quantities is to start with finding an *integer* value of D_{12} which maximizes the following likelihood function.

$$L(D_{12} | B_1, B_2, B_{12}) = \binom{D_{12} + 1}{B_{12}} \binom{B_1 + B_2 + 2 B_{12} - 2 D_{12} - 3}{D_{12} - B_{12}} \binom{B_1 + B_2 + 2 B_{12} - D_{12} - 2}{D_{12}}^{-1}. \quad (9)$$

The values of D_1 and D_2 are then found from the following equation

$$D_i = B_i + B_{12} - D_{12} - 1 \quad (i = 1, 2). \quad (10)$$

Expressing N_{12} in terms of D_1 , D_2 , and D_{12} . Using the definition of conditional probability, we can express the expected values of D_1 , D_2 , and D_{12} in terms of the probabilities of observing configurations C_i 's and K_j 's as follows.

$$E(D_i) = P_{CL} N \sum_{j=1}^8 \text{Prob}\{C_i | K_j\} \text{Prob}\{K_j\}, \quad (11)$$

where C_i 's and K_j 's are configurations defined in Equations 1 and 2, N is the common length of sequences S_1 , S_2 , and S_r , and i takes values from the set $\{1, 2, 12\}$.

In order to find an explicit form of the system 11, we need to derive probabilities $\text{Prob}\{C_i | K_j\}$'s and $\text{Prob}\{K_j\}$'s. Since many of the $\text{Prob}\{C_i | K_j\}$'s are equal with each other, it is convenient to show them as entries of a (3×8) matrix, \mathbf{M} , where

$$M_{1i} = \text{Prob}\{C_1 | K_i\}, \quad M_{2i} = \text{Prob}\{C_2 | K_i\}, \quad \text{and} \quad M_{3i} = \text{Prob}\{C_{12} | K_i\}. \quad (12)$$

After a few algebraic operations we find that

$$\mathbf{M} = \begin{pmatrix} 0 & \alpha & 0 & 1 & 0 & 0 & \gamma & 0 \\ 0 & \alpha & 0 & 0 & 1 & \gamma & 0 & 0 \\ 0 & \beta & 1 & 0 & 0 & \delta & \delta & 1 \end{pmatrix}, \quad (13)$$

where

$$\alpha = \sum_{i=1}^4 \pi_i \sum_{j \neq i} \frac{\pi_j^2 (1 - \pi_i - \pi_j)}{(1 - \pi_i)^3}, \quad \beta = \sum_{i=1}^4 \pi_i \sum_{j \neq i} \frac{\pi_j (1 - \pi_i - \pi_j)^2}{(1 - \pi_i)^3},$$

$$\gamma = \sum_{i=1}^4 \pi_i \sum_{j \neq i} \frac{\pi_j^2}{(1 - \pi_i)^2}, \quad \delta = \sum_{i=1}^4 \pi_i \sum_{j \neq i} \frac{\pi_j (1 - \pi_i - \pi_j)}{(1 - \pi_i)^2}.$$

and

(14)

In the absence of specific information about the equilibrium nucleotide frequencies (π_i 's), we can assume that $\pi_i = 1/4$ for all i , in which case constants α , β , γ , and δ assume values $2/9$, $4/9$, $1/3$, and $2/3$, respectively.

Noting that each of the eight configurations, K_1 , ..., and K_8 , represents an outcome of three independent random events, we can express $\text{Prob}\{K_j\}$'s in terms of parameters p_1 , p_2 , and p_r . After substituting these probabilities into Equation 11 and combining the result with Equation 13, we obtain the following system.

$$E(D_1) = NP_{CL} (\alpha p_1 p_2 p_r + p_1 q_2 q_r + \gamma q_1 p_2 p_r),$$

$$E(D_2) = NP_{CL} (\alpha p_1 p_2 p_r + q_1 p_2 q_r + \gamma p_1 q_2 p_r), \quad \text{and}$$
(15)

$$E(D_{12}) = NP_{CL} (\beta p_1 p_2 p_r + q_1 q_2 p_r + \delta p_1 q_2 p_r + \delta q_1 p_2 p_r + p_1 p_2 q_r),$$

where (where $q_j = 1 - p_j$). In the general case, an analytical solution with respect to p_1, p_2 , and p_r for this system of Equations could not be found, and one has to use one of the numerical algorithms. We solved the system 15 by searching for the values of parameters p_1, p_2 , and p_{12} that minimize the following residual sum of squares.

$$\left[\hat{D}_1 - f_1\right]^2 + \left[\hat{D}_2 - f_2\right]^2 + \left[\hat{D}_{12} - f_{12}\right]^2, \quad (16)$$

where \hat{D}_i 's are estimates of $E(D_i)$'s, and f_1, f_2 , and f_{12} are the right-hand sides of the three equations in the system 15, respectively. Minimization was done with a quasi-Newton algorithm implemented in a computer subroutine kindly provided to us by Dr. Ziheng Yang.

Once the estimates of p_1 and p_2 (\hat{p}_1 and \hat{p}_2 , respectively) are obtained, they can be used to compute an estimate of N_{12} (the number of differences between sequences S_1 and S_2) with the following Equation by substituting parameters in the right-hand side with their estimates.

$$N_{12} = N(p_1 + p_2 - p_1 p_2 \text{Prob}\{S_{1i} = S_{2i} | C_{12}\}). \quad (17)$$

Incidentally, we have $\text{Prob}\{S_{1i} = S_{2i} | C_{12}\} = \gamma$, where γ is as given in Equation 14.

In the case when p_1, p_2 , and p_r are small, the following *approximate* expressions are useful.

$$\begin{aligned} \hat{N}_{12} &\approx N(\hat{p}_1 + \hat{p}_2) \\ &= (\sqrt{8B_{12} + 8B_1 + 1} + \sqrt{8B_{12} + 8B_2 + 1} - 2\sqrt{8B_{12} + 1}) / (2PCL), \\ &\quad \text{(if digestion is partial),} \quad (18) \\ &= (3B_{12} + B_1 + B_2 - 1 - \sqrt{9B_{12}^2 + 4B_{12}\{B_1 + B_2\} - 6B_{12} + 1}) / P_{CL}, \\ &\quad \text{(if digestion is total).} \end{aligned}$$

The actual digestion pattern may turn out to be something in between the total and partial

digestions; then the formulas for computing N_{I2} under “ideal” partial and total digestions can be viewed as upper- and the lower-bound estimates.

In the case of the partial digestion the variance of the estimated distance can be approximately calculated with expression

$$\hat{V}(\hat{N}_{I2}) = (\sqrt{8B_{I2} + 8B_1 + 1} + \sqrt{8B_{I2} + 8B_2 + 1} - 2\sqrt{8B_{I2} + 1})_{(1 - PCL)/2}, \quad (19)$$

while an estimate of $V(\hat{N}_{I2})$ for the total digestion can be computed in the following way.

$$\begin{aligned} \hat{V}(\hat{N}_{I2}) = & (3B_{I2} + B_1 + B_2 - 1 - \sqrt{9B_{I2}^2 + 4B_{I2}\{B_1 + B_2\} - 6B_{I2} + 1})_{(1 - PCL)} \\ & + 4\hat{V}(\hat{B}_{I2})(b_1 + b_2 - b_{I2})^2 / (P_{CL})^2, \end{aligned} \quad (20)$$

where

$$\hat{V}(\hat{B}_{I2}) = \frac{\hat{D}_{I2}(\hat{D}_{I2} + 1)(\hat{D}_1 + \hat{D}_2 - 1)(\hat{D}_1 + \hat{D}_2)}{(\hat{D}_1 + \hat{D}_2 + \hat{D}_{I2} - 1)(\hat{D}_1 + \hat{D}_2 + \hat{D}_{I2})^2}, \quad (21)$$

$$b_1 = 1 - B_{I2}/S, \quad b_2 = -B_{I2}/S, \quad b_{I2} = [3 - \{9B_{I2} + 2(B_1 + B_2) - 3\}] / (2S),$$

$$S = \sqrt{9B_{I2}^2 + 4B_{I2}\{B_1 + B_2\} - 6B_{I2} + 1},$$

and \hat{D}_i 's are computed by substituting $E(B_j)$'s in the right-hand sides of Equations 8 with their

observed values, B_j 's. Both formulas for estimating $V(\hat{N}_{I2})$ were obtained by applying the “delta-technique” and the well-known statistical identity $V(X) = V[E(X|Y)] + E[V(X|Y)]$.

Computer simulation. The simulation scheme was very similar to those described by

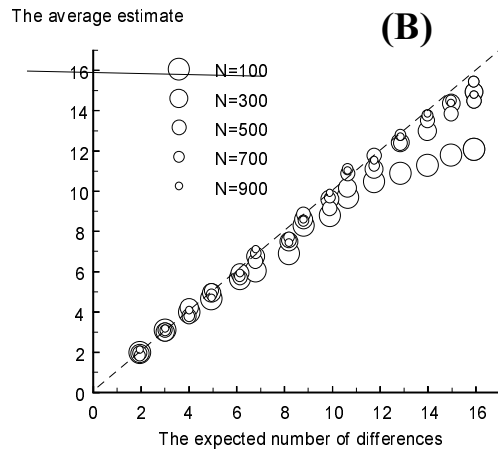
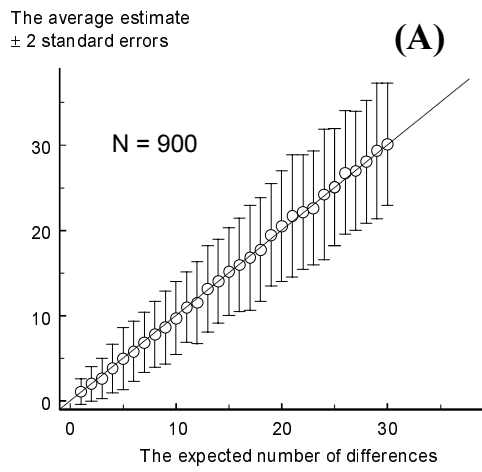


Figure 2.

DOPAZO, SOBRINO *et al.* (1993). First, an “ancestral” sequence was generated as a random sequence of equiprobable nucleotides. Second, the ancestral sequence was duplicated to give rise to the “extant” sequences S_r , S_l and S_2 . Third, each of the three “extant” sequences was randomly modified at each site with probabilities p_r , p_l and p_2 , respectively. Fourth, hybridization between S_r and S_l , and S_r and S_2 , and the digestion by RNase A were modelled. [In the simulations, we used the probabilities of cleavage for each type of mismatch that were estimated by PERUCHO (1989), so that $P_{CL} \approx 0.45$.

Although we simulated both partial and total digestion schemes, only the results of the former are shown here (Figure 2).] Fifth, the numbers of shared and non-shared fragments, B_1 , B_2 , and B_{12} were computed. We computed these numbers under two set-ups: (i) when *all* assumptions of our model are satisfied, so that the “observed” values of B_1 , B_2 , and B_{12} were equal to their actual values (Figure 2 A), and (ii) assuming that different heteroduplex fragments of the same size were lumped into a single electrophoretic band, so that the “observed” values of B_1 , B_2 , and B_{12} were *smaller* than their actual values (Figure 2 B). Finally, the “observed” values of B_1 , B_2 , and B_{12} were used to estimate N_{12} with Equations 4, 8, 15, and 17, as described above.

We simulated RAMCM with five different riboprobe sizes, 100, 300, 500, 700 and 900 base pairs, with 200 simulation replications per each set of parameter values. In all simulations required for generating plots in Figure 2 we used the following parameter values: $P_{CL} = 0.45$, $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$, $p_r = 0$, $p_l = (1 + i)/N$, $p_{12} = (1 + i)/N$, $p_2 = (p_{12} - p_l)/(1 - \gamma p_l)$, where $i = 1, 2, \dots, 30$ for Figure 2A, and $i = 1, 2, \dots, 16$, for Figure 1B. The estimation of N_{12} was done by numerically solving the system of Equations 15.

Results of simulation. The simulations strongly indicated that the estimator of N_{12} has the desired statistical properties when the model assumptions are fully satisfied. Figure 2A shows the typical result of computer simulation in which all assumptions are satisfied ($N=900$). The plots computed for shorter or longer riboprobes are virtually identical to the plot shown in Figure 2A, because the correspondence between the average estimate and the expected number of difference always remains strong, and the variance of the estimate is determined by the value of P_{CL} (the variance is large for small values of P_{CL}) rather than by the length of the riboprobe. As is expected, an increase of any combination of the values of p_1 , p_2 , and p_r also results in the growth of the variance of the distance estimate. Therefore, the best choice of the riboprobe for a set of genes is an “average” sequence which is close to the nearest common ancestor of all sequences in the set.

Application of the approximate Equation 17 instead of numerically solving system 15 produced more erratic but very similar results (data not shown), provided that the expected distance between S_1 and S_2 was not too large. Therefore, Equation 17 appeared to be sufficiently accurate for the practical calculations.

Another series of simulations (Figure 2B) showed that the algorithm is robust to violation of certain assumptions when the divergence between sequences S_1 and S_2 is not large ($N_{12} \leq 15$). More specifically, in the second series of simulations we assumed that non-identical fragments of heterohybrid molecules having the same size form single “electrophoretic bands” and can not be distinguished. (This assumption is unfavorable for our estimation procedure and represents “the worst case” in an experiment because fragments with the *same size* but *different nucleotide sequences* usually have *different* mobilities in the electric field.) As the result, the average distance estimate tends to be smaller

than the “true” value (see Figure 2B), although for small distances the bias is negligible. The standard errors of estimates were very close to those in Figure 2 A and are not shown on the plot.

Computer simulation clearly indicated that for the best results of data analysis, the length of the riboprobe should be selected in such a way that the distance between sequences under comparison does not exceed 15 differences. This can be accomplished by substituting one long riboprobe with a set of shorter non-overlapping probes. The overall distance between two long sequences should be then computed as a sum of distances estimated with each riboprobe.

To illustrate the application of the suggested algorithm to real data, we analysed the proximity of gII protein genes of several strains of the pseudorabies virus.

ASSESSING PROXIMITY OF VIRAL STRAINS

Gene gII of pseudorabies virus. Pseudorabies virus (PRV) belongs to subfamily alpha-herpesviridae of the herpesviridae family of RNA viruses. It is known to cause severe damages to the nervous system of affected swine; a complex of symptoms caused by PRV infection was described as the Aujeszky disease. For our analysis we have chosen PRV gene gII encoding a glycoprotein that appears on the surface of a mature virus particle and is recognized by the host immune system.

Viral DNA preparation. We used nine PRV strains derived from field isolates. The strains were obtained from a few different sources: from National Veterinary Services Laboratory in Ames, Iowa (strains Powlen, BE71, Indiana/Funkhauser, and Shope), from Dr. Maes at the Michigan State University, East Lansing (strain P2208), from the Ameri-

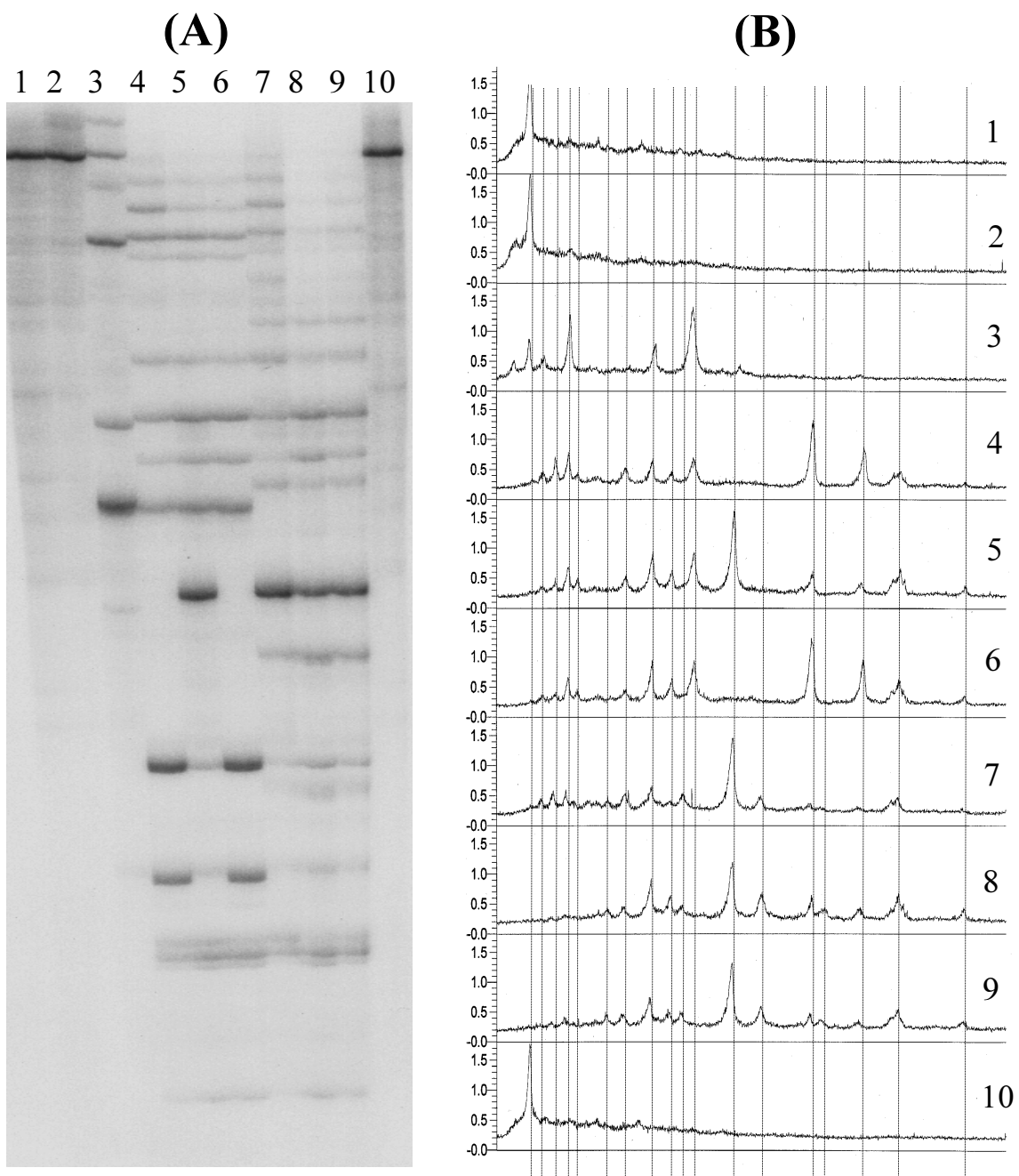


Figure 3.

2	0									
3	3.8	3.8								
4	8.1	8.1	4.3							
5	8.5	8.5	4.7	0.5						
6	8.1	8.1	4.3	0	0.4					
7	8.9	8.9	6.6	3.6	3.1	3.6				
8	8.1	8.1	9.3	4.9	4.2	4.9	2.6			
9	8.1	8.1	9.3	4.9	4.2	4.9	2.6	0		
10	0	0	3.8	8.1	8.5	8.1	8.9	8.1	8.1	
	1	2	3	4	5	6	7	8	9	

(A)

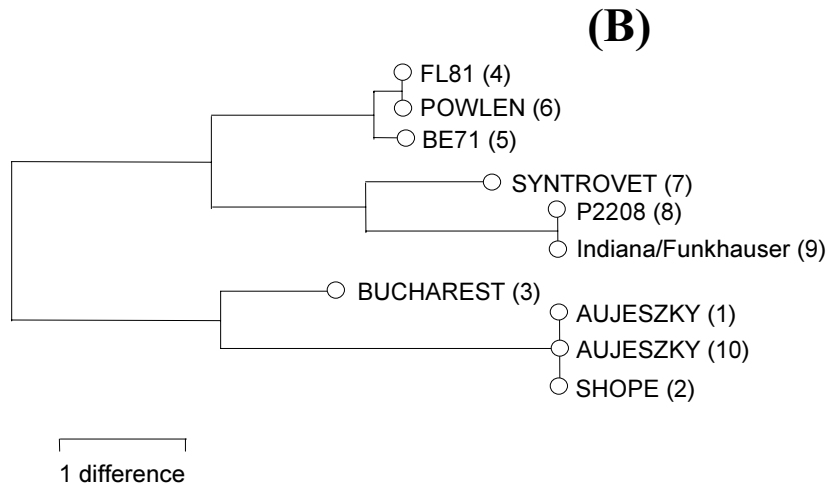


Figure 4.

can Type Culture Collection (strains Aujeszky and Bucharest), and from Dr. Hahn at the University of Illinois, Urbana-Champaign (strain FL81). Strains P2208 and Indiana/Funkhauser are viral clones from the same field strain, and served in our experiment as an index of possible intrastrain genetic drift. We propagated these eight PRV strains *in vitro* along with a commercial vaccine strain derived from the Iowa strain, and prepared samples of viral DNA for subsequent amplification [see DANGLER *et al.* (1992) for details].

Primers. We selected a pair of primers for PCR amplification from a published sequence for the PRV gII gene, using a commercial software package (LOWE *et al.* 1990). The sense primer (gIISN) had the PRV specific sequence 5'CTTCAAGGAGAACATCGCCCC3', while the antisense primer (gIIASN) had sequence 5'ACGTGCGTGCTGTTGTAGCG3'. We also synthesized another set of primers in which the SP6 RNA polymerase promoter sequence, 5'ATTTAGGTGACACTATAGAA3', was concatenated to the 5' end of each virus-specific primer sequence (SP6/gIISN and SP6/gIIASN).

Polymerase chain reaction. We amplified viral DNA in parallel amplification reactions using as a negative control DNA from Vero cells prepared as described by DANGLER *et al.* (1992). (In amplification of the PRV gII gene, each 100 μ l reaction contained 3 μ l of viral DNA suspension, 6% DMSO, 3 units of Taq DNA polymerase, 0.75 μ M MgCl₂, and 400 nM of each primer.) Two different primers were used for amplification of each strain: (i) SP6/gIISN and gIIASN, and (ii) gIISN and SP6/gIIASN. Reactions were run in a programmable heating block for 35 cycles with the following temperature profile: cycles 1 to 34, 95°C for 30 seconds, 50°C for 30 seconds, and 72°C for 45 seconds; cycle 35, 95°C for 30 seconds, 50°C for 30 seconds, and 72°C for 5 minutes. The negative control Vero

cell DNA consistently yielded negative amplification results.

RNA mismatch cleavage. We amplified viral RNA samples using the PCR products as transcription templates and following specifications provided by the manufacturer of a commercial kit for RNA transcription (Epicentre Technologies, Madison, WI). The presence of the SP6 polymerase promoter sequence on the 5' end of either the sense or antisense strand permitted the synthesis of complementary RNA strands; the sense RNA strand was radiolabelled with ^{32}P during the transcription reaction.

Next, we performed a series of hybridizations in which the radiolabelled sense RNA strand from the Aujeszky strain was used as the probe to the antisense RNA strand from each PRV strain. For the RAMCM steps of hybridization and enzymatic cleavage we used reagents from a commercial RAMCM kit (Ambion, Austin, TX) following manufacturer's instructions. (We used in this experiment the proprietary mismatch RNase stock solution #2, and the cleavage reactions were incubated at 70°C to reduce intramolecular interactions.) The digested RNA mixtures were analyzed by electrophoresis through a 6% polyacrylamide sequencing gel run at 50°C and gel autoradiography (see Figure 3).

Analysis of RAMCM patterns. We analysed the resulting RAMCM patterns (Figure 3A) with an automatic densitometer (see Figure 3B) and then converted them into a matrix of differences between strains (Figure 4A) using the approximate formula (17). The distance matrix was then used to compute a dendrogram (Figure 4B) with the neighbor-joining method (SAITOU and NEI 1987) by the program METREE (Rzhetsky and Nei 1994). Note that in this example the maximum value in the distance matrix does not exceed 9 (see fig. 4 A), so that the estimates should be virtually unbiased. The gene “genealogy” obtained from RAMCM patterns is in perfect correspondence with the rela-

tionships among the PRV strains proposed from independent considerations (DANGLER *et al.* 1992).

DISCUSSION

Properties of the proposed method. RAMCM is an inexpensive, sensitive, and fast procedure for detecting differences between related sequences (ERLICH and ARNHEIM, 1992), which is now applicable to quantitative analyses. Previously, this method was relegated to qualitative use, in which evaluation of dissimilarity between sequences was both cumbersome and subjective, seriously limiting the overall utility of the approach.

The performance of the procedure could be substantially improved by fine-tuning the experimental stage of RAMCM. Indeed, our computer simulation indicated that the largest contribution to the overall error of estimated differences comes from the existence of mismatches that are “undetectable” by RNase A. Therefore, statistical efficiency of the estimates could be significantly improved by modifying the experimental procedure aiming to decrease the proportion of “undetectable” mismatches.

As we have shown in computer simulations, the RAMCM technique gives the best results when applied to closely related sequences, up to a maximum of 15 differences. This is because, as the number of mismatches increases, the overall number of electrophoretic bands grows explosively under the partial digestion conditions, causing errors in the counting shared and unique bands between electrophoretic lanes. If the sequences under analysis are likely to be separated by large distances, we recommend substituting a single long riboprobe (which is commonly used in the studies of this kind, see LÓPEZ-GALÍNDEZ, LÓPEZ *et al.* 1988; ROJAS, DOPAZO *et al.* 1995; OWEN and PAULUKAITIS, 1988; GARCÍA, MARTÍN *et al.* 1994; LÓPEZ-GALÍNDEZ, ROJAS *et al.* 1991) with a set of

shorter non-overlapping probes, selected to maximise the precision of estimation.

Further improvement of the model. In the future, one may attempt to develop a model accounting for additional features of the real nucleotide genes and peculiarities of RAMCM technique. For example, in certain situations it may prove useful to take into account the following aspects: (1) in some sequences, the distribution of mutations is significantly non-random, which is usually manifested as an existence of “conservative” and “mutation hotspot” sites, (2) the actual causes of underestimation of the observed number of electrophoretic bands are sometimes known and can be incorporated into the model, and (3) in some sequences insertions and deletions are not infrequent and should be included into the model. Although the model suggested in our study is somewhat oversimplified, the predictions it provides seem to be sufficiently accurate to make the inclusion of many of the additional factors unnecessary in most actual data analyses.

Conclusion. We proposed an algorithm that opens possibility of using RAMCM as a quantitative tool in large-scale sequence analyses. Although the RAMCM does not provide as much information as direct nucleic acid sequencing, it allows for quick and inexpensive examination of a large number of samples and constitutes a feasible alternative for population genetics and molecular epidemiology studies.

LITERATURE CITED

- AARONSON, R.P., YOUNG, J.F. and PALESE, P., 1982. Oligonucleotide mapping: evaluation of its sensitivity by computer simulation. *Nucl. Acids Res.* **10**:237-246.
- CRISTINA, J., MOYA, A., ARBIZA, J., RUSS, J., HORTAL, M., ALBÓ, C., GARCÍA-BARRENO, B., GARCÍA, O., MELERO, J.A. and PORTELA, A., 1991. Evolution of the G and P genes of human respiratory syncytial virus (subgroup A) studied by the RNase A mismatch cleavage method. *Virology* **184**: 210-218.
- DANGLER, C.A., DEAVER, R.E., KOLODZIEJ, C.M., and RUPPRECHT, J.D., 1992. Genotypic screening of pseudorabies virus strains for thymidine kinase deletions by use of the polymerase chain reaction. *Am. J. Vet. Res.* **53**: 904-908.
- DOPAZO, J., SOBRINO, F. and LÓPEZ-GALÍNDEZ, C., 1993. Estimates by computer simulation of genetic distances from comparison of RNase A mismatch cleavage patterns. *J. Virol. Meth.* **45**: 73-82.
- ERLICH, H.A. and ARNHEIM, N., 1992. Genetic analysis using the polymerase chain reaction. *Annu. Rev. Genet.* **26**: 479-506.
- FORRESTER, K., ALMOGUERA, C., HAN, K., GRIZZLE, W.E. and PERUCHO, M., 1987. Detection of high incidence of *K-ras* oncogenes during human colon tumorigenesis. *Nature* **327**: 298-303.
- GARCÍA, O., MARTÍN, M., DOPAZO, J., ARBIZA, J., FRABASILE, S., RUSSI, J., HORTAL, M., PÉREZ-BREÑA, P., MARTÍNEZ, I., GARCÍA-BARRENO, B., *et al.*, 1994. Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G protein. *J. Virol.* **68**: 5448-5459.

JUKES, T. H, CANTOR, C. R., 1969. *in*: Munroe, H. L., editor. Mammalian protein metabolism. New York: Academic Press, *Evolution of protein molecules*. p. 21-132.

KEILSON, J., 1979. *Markov chain models - rarity and exponentiality*. Springer-Verlag, New York.

LÓPEZ-GALÍNDEZ, C., LÓPEZ, J.A., MELERO, J.A., DE LA FUENTE, L., MARTÍNEZ, C., ORTÍN, J. and PERUCHO, M., 1988. Analysis of genetic variability and mapping of mutations in influenza virus by the RNase A mismatch cleavage method. *Proc. Natl. Acad. Sci. USA* **85**: 3522-3526.

LÓPEZ-GALÍNDEZ, C., ROJAS, J.M., NÁJERA, R., RICHMAN, R.R. and PERUCHO, M., 1991. Characterization of genetic variation and 3'-azido-3'deoxythymidine-resistance mutations of human immunodeficiency virus by the RNase A mismatch cleavage method. *Proc. Natl. Acad. Sci. USA* **88**: 4280-4284.

LÓPEZ-GALÍNDEZ, C., ROJAS, J.M., and DOPAZO, J., 1995. The RNase A mismatch method for the genetic characterization of viruses, pp. 547-552 in *Molecular basis of virus evolution*, edited by A. Gibbs, G.H. Calisher, and F. Garcia-Arenal. Cambridge University Press, Cambridge.

LOWE, T., SHAREFKIN J., and YANG, S.Q., 1990. A computer program for selection of oligonucleotide primers for polymerase chain reactions. *Nuc. Acids Res.* **18**: 1757-1761.

MYERS, R.M., LARIN, Z. and MANIATIS, T., 1985. Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. *Science* **230**:1242-1246.

NEI, M., 1987. *Molecular evolutionary genetics*. Columbia University Press. New York.

- NEI, M., and JIN, L., 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* **6**: 240-300.
- NEI, M. and LI, W., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**: 5269-5273.
- OWEN, J. and PAULUKAITIS, P., 1988. Characterization of cucumber mosaic virus I. Molecular heterogeneity mapping of RNA 3 in eight CMV strains. *Virology* **166**:495-502.
- PERUCHO, M., 1989. Detection of single-base substitutions with the RNase A mismatch cleavage method. *Strat. Mol. Biol.* **2**: 37-41.
- ROJAS, J.M., DOPAZO, J., SANTANA, M., LÓPEZ-GALÍNDEZ, C. and TABARÉS, E., 1995. Comparative study of the genetic variability in thymidine kinase and glycoprotein B genes of herpes simplex viruses by the RNase A mismatch cleavage method. *Virus. Res.* **35**: 205-214.
- RZHETSKY, A., and NEI, M., 1994. METREE: Program package for inferring and testing minimum-evolution trees. *Comp. Appl. Biol. Scien.* **10**: 189-191.
- SAITOU, N., and NEI, M., 1987. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.
- SÁNCHEZ-PALOMINO, S., DOPAZO, J., OLIVARES, I., MARTIN, M.J., and LÓPEZ-GALÍNDEZ, C., 1995 Primary genetic characterization of HIV-1 isolates from WHO-sponsored vaccine evaluation sites by the RNase A mismatch method. *Viruses Res.* **39**: 251-259.
- WHO NETWORK FOR HIV ISOLATION AND CHARACTERIZATION, 1994. HIV type 1 variation in world health organization-sponsored vaccine evaluation sites: genetic screen-

ing, sequence analysis, and preliminary biological characterization of selected viral strains. *AIDS Research and Human Retrovirus* **10**: 1327-1343.

WINTER, E., YAMAMOTO, F., ALMOGUERA, C., and PERUCHO, M., 1985. A method to detect and characterize point mutations in transcribed genes: amplification and overexpression of the mutant c-Ki-ras allele in human tumor cells. *Proc. Natl. Acad. Sci. USA* **82**: 7575-7579.

FIGURE LEGENDS

Figure 1. (A) An hypothetical unrooted tree relating the three “extant” sequences, S_1 , S_2 , and S_r . In the model described in the text, S_1 and S_2 are the sequences to be compared and S_r is a reference sequence used as a probe; S_a is an hypothetical sequence which is ancestral to sequences S_1 , S_2 , and S_r . Parameters p_1 , p_2 , and p_r specify the expected proportions of differences between the “ancestral” sequence S_a and the “extant” sequences S_1 , S_2 , and S_r , respectively. (B) Schematic representation of heterohybrid molecules: the reference sequence is shown in black, and sequences S_1 and S_2 are shown as gray strands in heterohybrids; white bulges indicate *detectable* mismatches, *i.e.*, the mismatches that are recognized by RNase A. D_i 's ($i = 1, 2, 12$) denote the numbers of *detectable* mismatches that are common ($i = 12$) and unique ($i = 1, 2$) for the two heterohybrid molecules. Figures (C) and (D) show “ideal” partial and total digestion patterns, respectively, observable in an hypothetical electrophoretic analysis (compare with Figure 1 B). In the *partial* digestion all possible subfragments resulting from partial cleavage of *detectable* mismatches are generated. In the *total* digestion, *all detectable* mismatches are cleaved in all heteroduplexes. Variables B_i 's denote the observed number of electrophoretic bands, where B_{12} stands for the number of band *common* for two electrophoretic lanes, and B_1 and B_2 are the numbers of bands *unique* for each line.

Figure 2. Results of computer simulation of RAMCM technique where (A) all model assumptions were satisfied, and (B) non-homologous fragments of the same length were assumed to be indistinguishable. Each data point in each plot was obtained by averaging 200 computer simulations. The sequences lengths $N = 100, 200, 300, 500, 700,$ and 900 were studied in both cases, although only $N = 900$ is presented in the former case (A) because the plots obtained for different sequence lengths were virtually identical. The standard errors (not shown in plot B) were virtually identical in all simulations.

Figure 3. A photograph of electrophoretic gel (A) and results of densitometric analysis of this gel (B) showing RAMCM patterns resulting from analysis of gII protein genes from several strains of pseudorabies virus. The numbers shown on the top of the electrophoretic gel correspond to the PRV strains in the following order: (1) and (10) - AUJESZKY strain, which was used as a probe to all other sequences, (2) - SHOPE, (3) - BUCHAREST, (4) - FL81, (5) - BE71, (6) - POWLEN, (7) - SYNTROVET, (8) - P2208, and (9) - Indiana/Funkhauser.

Figure 4. (A) The matrix of distances between genes expressed in terms of the raw number of differences computed from analysis of RAMCM patterns in Figure 3. (B) The neighbor-joining tree (algorithm by Saitou and Nei 1987; implemented in program METREE by Rzhetsky and Nei 1994) computed from this matrix. The tree is in a good correspondence with independent data on the relationship among these PRV strains.

