Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10320–10324, November 1993
Evolution

Split decomposition: A technique to analyze viral evolution

(evolutionary trees/quasispecies/overlapping clustering)

JOAQUIN DOPAZO*†, ANDREAS DRESS‡, AND ARNDT VON HAESELER‡§

*Department of Genetics, University of Valencia, E-46100 Burjassot (Valencia), Spain; and ‡Department of Mathematics, University of Bielefeld, W-4800 Bielefeld, Germany

Communicated by Gian-Carlo Rota, June 4, 1993

ABSTRACT **A clustering technique allowing a restricted amount of overlapping and based on an abstract theory of coherent decompositions of finite metrics is used to analyze the evolution of foot-and-mouth disease viruses. The emerging picture is compatible with the existence of viral populations with a quasispecies structure and illustrates various forms of evolution of this virus family. In addition, it allows the correlation of these forms with geographic occurrence.**

Sequence data obtained in the last few years raise various interesting and important problems concerning *viral evolution*. In particular, viruses do not always evolve in a definite tree-like manner, as most higher organisms do. Instead, it has been found (1, 2) that many RNA virus populations exhibit characteristics typical of a *quasispecies* (3, 4).

Therefore it does not seem appropriate in this context to apply any one of the available tree (re)construction methods (5–9), which approximate data by tree-like structures only. Instead, we propose to use a *nonapproximative* method, developed in ref. 10.

This method associates to every finite metric (or dissimilarity measure) d, defined on a finite collection X of objects, a family of weighted *splits* $S{:}X = A \dot{\cup} B$ of X into a pair of nonempty, disjoint subsets A, B. Apart from a well-defined undecomposable residual constituent d_0 of d, which represents the unresolvable *noise* in the data, the metric d can be reconstructed from the associated family of splits simply by superposition. The splits suggest how a given collection of data (e.g., viruses) may naturally be divided into subfamilies. Their significance can be checked by Monte Carlo experiments.

We will exemplify this method by applying it to the gene encoding the VP1 capsid protein, which carries most of the responsibility for the antigenic reactivity of the foot-and-mouth disease (FMD) virus. FMD virus is a picornavirus, causing an economically serious viral disease in cattle and other cloven hooved animals (11). It is classified into seven immunologically distinguishable types: A, C, O, SAT1, SAT2, SAT3, and Asia 1. Although it has been shown (12) that the relation among the A, C, and O types is more or less tree-like, there is much evidence that viral populations also show features typical of a quasispecies (12–14), which may cause serious drawbacks when searching for phylogenetic relationships. In particular, an open question is the kinship relation within each A, C, and O type, which could not be resolved in ref. 12. We will demonstrate that the split decomposition technique elucidates these kinship relations remarkably well. While certain splits correlate very well with the geographic distribution of the viruses, their overall distribution also allows one to distinguish different modes of evolution in A, C, and O types.

METHODS

Splits Associated with a Finite Metric. Let $X = \{s_1, \ldots, s_n\}$ denote a set of $n = \# X$ aligned sequences with a distance measure $d(s_\nu, s_\mu) \geq 0$. A pair of two nonempty, disjoint subsets A, B of X, with $A \cup B = X$ is called a d split (10) if the associated (and by definition nonnegative!) *isolation index*

$$\alpha_{A,B} = \alpha^d_{A,B} := \frac{1}{2}\min\left\{\max\left\{\begin{matrix} d(a,b) + d(a',b'),\\ d(a,b') + d(a',b),\\ d(a,a') + d(b,b') \end{matrix}\right\} - d(a,a') - d(b,b') \,\middle|\, \begin{matrix} a, a' \in A;\\ b, b' \in B \end{matrix}\right\}$$

is positive (see Fig. 1). Every metric d decomposes canonically into a (weighted) sum $d_s := \Sigma_{A,B}\alpha_{A,B}\cdot\delta_{A,B}$ of *split metrics* $\delta_{A,B}$, defined by $\delta_{A,B}(s_\nu, s_\mu) := 0$ if $s_\nu, s_\mu \in A$ or $s_\nu, s_\mu \in B$, and $:= 1$ otherwise, and a *split-prime* residue d_0 $(:= d - d_s)$, that is, a metric d_0 without any d_0 split. In Fig. 1, d_0 vanishes identically. In real data sets this is not always the case, while it holds for almost trivial reasons for every tree-like metric (10, 15).

The Data. The following sequences have been studied.¶

A serotype: A5Mor/83 (M16090), A5Port/83 (M16092), A5Sp/86 (16), A5Ww/51 (K03343), A5Bb/84 (M16083), A12/32 (M10975), A22/65 (K03343), A10/61 (M20715), A24/55 (J02183), A27/76 (K03341), A32/70 (K03342), AArg/79 (K03345), AVen/76 (K03344).

C serotype: C1Obb/60 (17), CS10/79 (M22502), CS15/81 (M22505), CS16/81 (M22506), CS20/80 (18), CS22/80 (18), CS30/82 (18), C3Res/55 (M19760), C3Ind/71 (K01202), C3Ind/78 (J02184), C3Arg/84 (M19761), C3Arg/85 (M19762).

O serotype: O1Bfs/68 (J02185), 01Ca/58 (M15978), O1Mu/82 (19), O2Norm/49 (M15986), OIsr/81 (M15985), OTh/81 (M15984), OWupp/82 (M15983).

The numbers in parentheses are the GenBank accession numbers. South American sequences are underlined. The number after the slash indicates the isolation year.

For three families of A, C, and O serotype FMD viruses, we have transformed the VP1 genes into sequences in the purine (R)/pyrimidine (Y) alphabet. This procedure reduces evolutionary noise, due to high transition probabilities (20). Subsequently, we calculated the Hamming distance, using for each serotype only those positions that were completely determined for every member of the family in question.

Abbreviation: FMD, foot-and-mouth disease.

†Present address: Departamento de Sanidad Animal, C.I.T.-Instituto Nacional de Investigaciones Agrarias, Calle Embajadores 68, E-28012 Madrid, Spain.

§Present address: Lehrstuhl f. allg. Biologie, Institut f. Zoologie, Luisenstrasse 14, W-8000 Munich 2, Germany.

¶The aligned sequences can be obtained from the authors upon request.

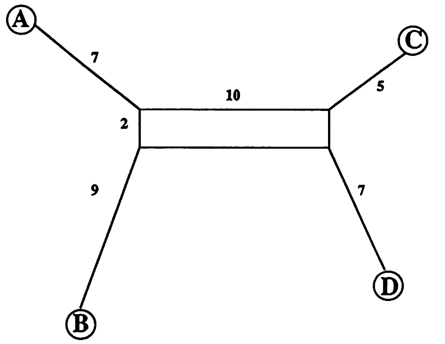

FIG. 1. Illustration of the distances among sequences A, B, C, and D. The pairwise distance is the sum of the shortest, weighted path connecting any two sequences in the graph. The weight of each edge is shown. The metric space can be decomposed in six splits with a positive isolation index: $\alpha_{\{A\},\{B,C,D\}} = 7$, $\alpha_{\{B\},\{A,C,D\}} = 9$, $\alpha_{\{C\},\{A,B,D\}} = 5$, $\alpha_{\{D\},\{A,B,C\}} = 7$, $\alpha_{\{A,B\},\{C,D\}} = 10$, $\alpha_{\{A,C\},\{B,D\}} = 2$. In this example the split-prime part of the metric equals 0.

For the resulting metric d we determined all splits and the split-prime residue d_0 as well as the parameters

$$D_0 := \sum_{\nu<\mu} d_0(s_\nu, s_\mu)$$

and

$$D_i := \sum_{\min(\#A,\#B)=i} \sum_{\nu<\mu} \alpha_{A,B} \cdot \delta_{A,B}(s_\nu, s_\mu)$$

$$= i \cdot (n-i) \cdot \sum_{\min(\#A,\#B)=i} \alpha_{A,B},\ i = 0, 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor.$$

Note that $\Sigma_{i=0}^{\lfloor \frac{n}{2} \rfloor} D_i = D := \Sigma_{\nu<\mu} d(s_\nu, s_\mu)$.

Monte Carlo Studies. To check the significance of our results, we interchanged the letters at each position of our aligned sequence families independently. This procedure leaves the consensus sequence and the sum of all distances invariant, but removes correlations within the sequences. For each set of randomized sequences of a given serotype, we calculated the parameters D_i ($i = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$) and determined their mean values $\overline{D}_i$.

RESULTS

Decomposition Results. Fig. 2 shows the decomposition of the metric spaces, given by the C and the O serotype FMD viruses. In both cases, we were able to draw a network that perfectly reflects the observed Hamming distances. In particular, the split-prime part d_0 turned out to be the so-called $K_{2,3}$ graph, which is shown in Fig. 2 *Inset*.

In Fig. 2*A*, describing the metric space spread out by the C-type sequences, we easily locate two well-separated sequence families. This separation is easily correlated with the geographic distribution of the virus. The first subgroup consists of seven European C1-subtype sequences; six of them, the CS group, belong to a single outbreak. The relation among these sequences is perfectly tree-like, probably because all CS types arose sequentially from a single outbreak.

The second group, the C3 subserotype, stemming from South America, shows a moderate degree of randomization, indicated by the non-tree-like distances. Evolution appears to follow two pathways in the C3 types.

In Fig. 2*B*, which displays the relation among O-type sequences, two subserotypes can be easily identified: O1

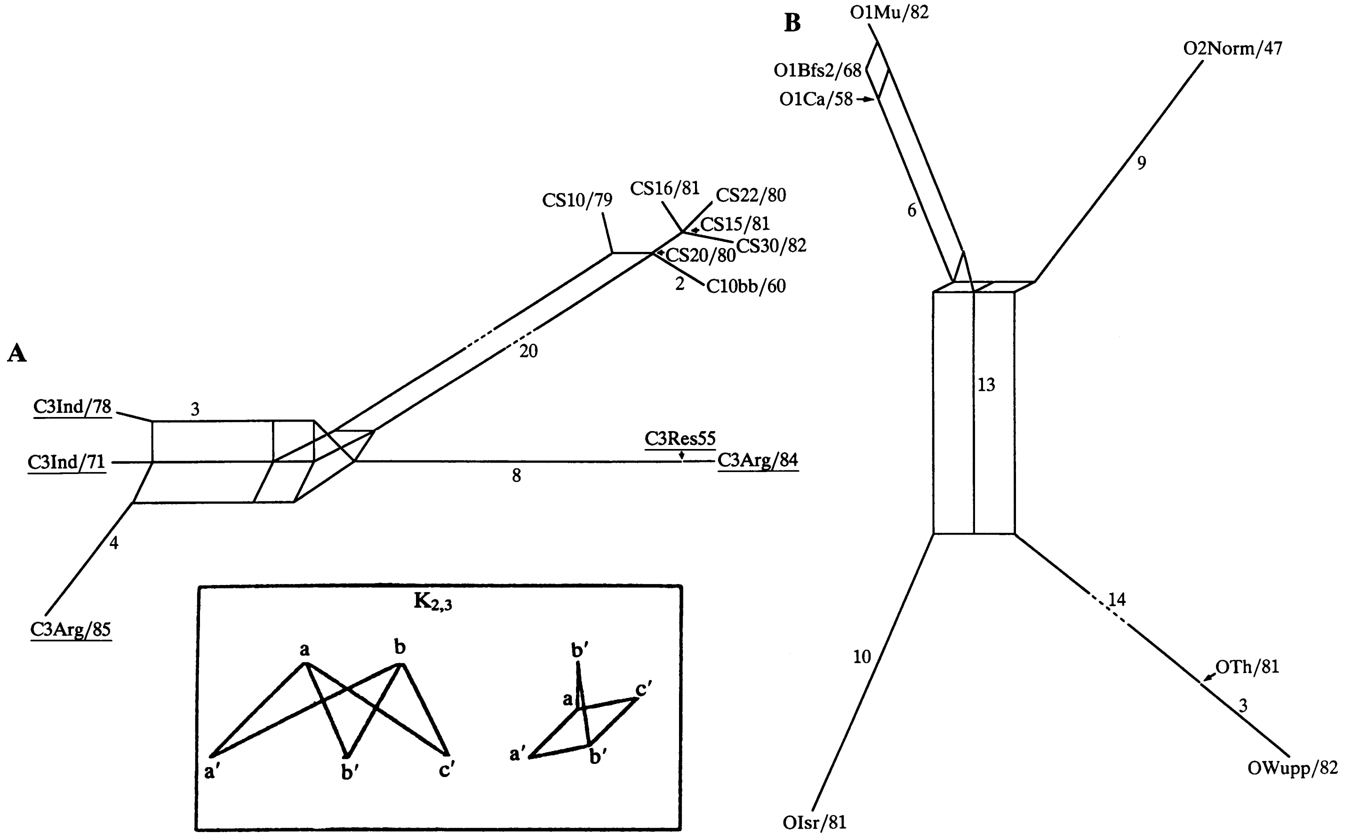

FIG. 2. Decomposition results of the Hamming metric the C type (*A*) and the O type (*B*) FMD virus VP1 genes. The split-prime part of both metrics is the graph $K_{2,3}$ (*Inset*). The minimal sum of weights of edges connecting two sequences in the network perfectly matches the Hamming distance between these sequences. The numbers along edges indicate the edge lengths. Edges without a number have unit length. South America viruses are underlined.

(composed of O1Mu/82, O1Bfs/68, and O1Ca/58) and O2 (represented by O2Norm/49). The remaining O-type sequences, yet unassigned to any subserotype, split into {OIsr/81} and {OTh/81 and OWupp/82}, both separated from the other O-type sequences by rather large isolation indices. They even may be of different origin (19), and they probably belong to different subserotypes. A moderate degree of randomization can be observed in the O type, too.

Unlike serotype C and O, the available data set of serotype A is composed of sequences belonging to many different subserotypes. Therefore, it is not astonishing that the d_0 part of the metric, extended to all A-type sequences, is too large and fuzzy to be drawn. Ergo, we restricted ourselves to the split decomposable part d_s, illustrated in Fig. 3*A*. This part looks very tree-like. Only two splits, both of isolation index 1, destroy the tree likeness of d_s. The asterisk in Fig. 3*A* indicates a small bunch of A5 sequences, which are distinguished only by the split-prime part d_0 and coincide relative to d_s. As we did not integrate the split-prime part d_0 into our illustration, some distances in the diagram appear to be smaller than they really are.

There exists a correspondence between the geographic place of isolation and the position occupied in the diagram. If the virus A27/76 from Colombia, showing the largest isolation index (and therefore having accumulated more *noisy* parallel mutations), is not taken into account, we can distin-

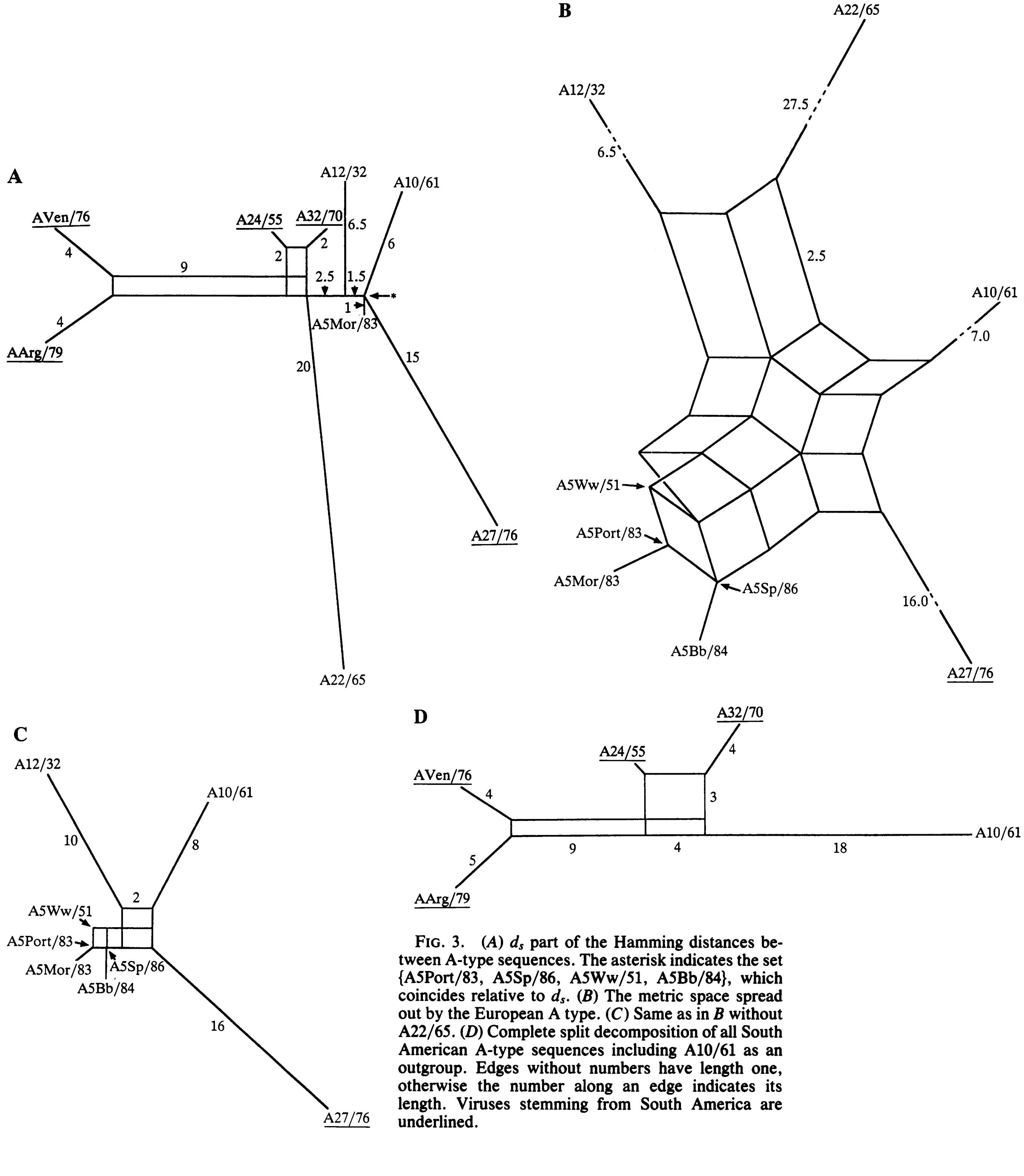

FIG. 3. (*A*) d_s part of the Hamming distances between A-type sequences. The asterisk indicates the set {A5Port/83, A5Sp/86, A5Ww/51, A5Bb/84}, which coincides relative to d_s. (*B*) The metric space spread out by the European A type. (*C*) Same as in *B* without A22/65. (*D*) Complete split decomposition of all South American A-type sequences including A10/61 as an outgroup. Edges without numbers have length one, otherwise the number along an edge indicates its length. Viruses stemming from South America are underlined.

guish a European group (the A5 group, A12/32, and A10/61) and a South America group (AArg/79, AVen/76, A32/70, and A24/55). We can determine different evolutionary patterns within the A sequences, which appear to be correlated with the geographic distribution of the sequence family. The South American sequences exhibit a high mutational rate and an almost tree-like evolution, while, as in the case of C types, the European A type is more conservative.

To resolve the European A type in more detail and independently of the evolutionary noise created by putting all A type sequences together, we studied also the split decomposition of the family, consisting of all A5 sequences together with A10/61, A12/32, A22/65, and A27/76. Now, the whole picture can be drawn (Fig. 3*B*). Furthermore, if we remove A22/65 from this set, the resulting metric is completely split decomposable and almost tree-like (Fig. 3*C*). The apparent, but small, nettedness of the diagram obtained indicates that the mode of evolution is that of a quasispecies. Finally, Fig. 3*D* displays the metric space of all South American A-type sequences and the A10/61 sequence. This space, which is completely decomposable, has a very tree-like structure, and the distances between the sequences are large.

Monte Carlo Experiments. To estimate the significance of our results, we carried out a number of Monte Carlo experiments, as described above.

At first, we asked which isolation indices for a fixed split size $i = \min(\#A, \#B)$ are significant. Second, we compared the relative proportion of each D_i.

To begin with, we discuss the mean isolation indices of a given split size. Table 1 shows for a split size i the mean isolation indices $\bar{\alpha}_O(i)$, $\bar{\alpha}_C(i)$, $\bar{\alpha}_A(i)$ and $\bar{\alpha}^r_O(i)$, $\bar{\alpha}^r_C(i)$, $\bar{\alpha}^r_A(i)$, respectively, for the randomized sequences. Except for the A type, the mean isolation indices for splits of size 1 increased in randomized sets, whereas isolation indices $\bar{\alpha}^r$ for splits of sizes greater than 1 are always considerably less than corresponding indices in the original data. Moreover, the number of splits with size >2 decreased remarkably. That is to say, the probability is very low to find a random split with size ≥3 and even less to find one with a high isolation index.

This contrasts with our real data. For example, in the C serotype, we have calculated a split of size 5 with isolation index of 20.00. In view of our randomization experiments, this split is highly significant (see Fig. 2*A*), whereas the split of size 4, separating C3Arg/85, C3Ind/71, C3Ind/78, and CS10/79 from the rest is not particularly significant. Discarding this split and ignoring the split-prime part of the metric, we obtain a tree that is in perfect agreement with *all* quartets as discussed in ref. 21.

Even for A-type sequences, we observe two splits of size 5 and 6, with isolation indices equal to 2.5 and 1.5, respectively. In the perturbed sets, not a single split of such size could be found. These results confirm the observed correspondence between geographic distribution and position in the diagram: by discarding the isolate A22/65 from USSR and the isolate A27/76 from Colombia because they show a large isolation index (more than twice the average index of splits of size 1 and greater than any average isolation index of any split-size; see Table 1) and because they probably are not too closely related to the other viruses in the group, the South America/Europe split has a significant isolation index of 2.5. The leftmost set in Fig. 3*A* is composed of South American viruses, and the other set is composed of A12/32, the A5 group, and A10/61, all European sequences.

Table 1. Mean isolation indices for different split sizes

Split size	C type		O type		A type	
(i)	$\bar{\alpha}_C$	$\bar{\alpha}^r_C$	$\bar{\alpha}_O$	$\bar{\alpha}^r_O$	$\bar{\alpha}_A$	$\bar{\alpha}^r_A$
1	1.44	4.62	5.75	9.09	6.61	6.62
2	8.00	1.20	7.50	1.66	5.50	1.01
3	3.00	1.06	6.67	1.36	1.00	0.79
4	1.00	1.00	—	—	—	0.75
5	20.00	1.00	—	—	2.50	—
6	—	—	—	—	1.50	—

Mean isolation indices for the undisturbed subtypes of the FMD virus ($\bar{\alpha}_C$, $\bar{\alpha}_O$, $\bar{\alpha}_A$) and mean isolation indices for randomized sequences ($\bar{\alpha}^r_C$, $\bar{\alpha}^r_O$, $\bar{\alpha}^r_A$) are given. To obtain a good estimate of these random values, almost 1000 Monte Carlo experiments were performed. However, one should keep in mind that the detection probability for large splits is very low; therefore, the average isolation indices may exhibit great fluctuations in the case of large splits.

In the O group, the largest observed split size is 3, but in this case the mean isolation indices for the splits are significantly larger than the randomized values. Only the split {O2Norm/49, OTh/81, OWupp/82}, {rest} does not seem to be significant.

Even more instructive is the contribution of the D_i values to the distance sum D. In Table 2 the results are listed: upon randomization $\bar{D}^r_0$ and $\bar{D}^r_1$ values increased in all three serotypes. The increase in the $\bar{D}^r_0$ value of the A type is clearly smaller than for the other types (e.g., 9-fold for the C type). The $\bar{D}^r_1$ value increases by a factor of 3–4, while for $i \geq 2$, the D_i values decrease drastically upon randomization.

We can summarize these results as follows. A metric, based on a set of randomized sequences, is mainly decomposed in splits of size 1, with a large mean isolation index and a large residue d_0. In contrast, real data like the C type have many larger splits and a small residue.

The A-type VP1 genes exhibit a somehow different behavior. As already mentioned, the split-prime part of the underlying metric is comparatively large. And there are no distinct differences between D_0 and $\bar{D}^r_0$ or D_1 and $\bar{D}^r_1$. The Monte Carlo study shows that the two large splits, with $\alpha = 2.5$ and 1.5, respectively, are significant and decompose the A type in three groups:

{AArg/79, AVen/76, A32/70, A24/55, A22/65},

{A12/32}, and

{A5Port/83, A5Sp/86, A5Mor/83, A5Ww/51, A5Bb/84,
A10/61, A27/76}.

One should keep in mind that in the A serotype a great amount of evolutionary noise, as expressed by the large D_0 value, has accumulated. Especially, most of the splits in Fig. 3 *B* and *C* are not significant. Longer sequences will certainly help to analyze the kinship relation of A types in more detail.

Table 2. Decomposition of the distance matrices

Split size	C type		O type		A type	
(i)	D_i	$\bar{D}^r_i$	D_i	$\bar{D}^r_i$	D_i	$\bar{D}^r_i$
0	61.0	574.7	26.0	110.4	422.0	542.0
1	143.0	608.0	138.0	381.8	714.0	1031.9
2	160.0	24.5	150.0	42.7	242.0	25.6
3	81.0	1.6	240.0	19.1	60.0	1.4
4	64.0	0.1	—	—	—	0.1
5	700.0	0.1	—	—	100.0	—
6	—	—	—	—	63.0	—

Absolute fraction of D_i values for the original sequence set and for the randomized sequences. In the latter case, the values shown are mean values from 1000 Monte Carlo experiments.

DISCUSSION

RNA virus populations consist of closely related but not identical genomes. Due to the limited fidelity of replication catalyzed by viral polymerase (10^{-3} to 10^{-5} mutations per site and replication cycle; see ref. 1 for a review), RNA viruses display extremely high rates of fixation of mutations. As a consequence of this, they form quasispecies (3, 4), consisting of a mutant spectrum centered around a "master" or "consensus" sequence. The evolution of such populations is an intricate dynamical process governed by the complex interplay between the host immune system and the virus and the selective pressure they exert on each other (as well as by vaccination policies). It involves the formation of (quasi-)stable equilibrium states as well as rapid emergence of new mutant spectra. So while there are episodes of clock-like accumulation of mutations (22), this pattern of evolution is not observed in general (12, 23–27). Accordingly, in (dis)similarity data derived from viral sequences, phylogenetic kinship relations can be masked by the quasispecies character of viral populations and by erratic mutation rates as well as by adaptive convergence and accumulation of chance events.

This motivates the use of an analytic tool called split decomposition, which searches not only for tree-like kinship relations but detects as well other forms of evolutionary dynamics.

As outlined in ref. 1, such a pool of mutants as found in quasispecies populations is able (*i*) to buffer changes, maintaining a consensus sequence in a stable environment for an undefined period; (*ii*) to gradually accumulate mutations under gradually changing selective pressure (as found during an epidemic), exhibiting a clock-like pattern of evolution; and (*iii*) to show huge changes in short periods of time when, during an outbreak and under severe rupture of the equilibrium, selection for rare variants present in the quasispecies is executed. If these ruptures are large enough, the arising quasispecies can constitute a new subserotype or even a new type.

All these features of evolutionary dynamics can be detected in the diagrams derived by split decomposition. The low correlation between isolation date and position in these diagrams combined with the clear correlation between position and geographic origin indicates that viruses belonging to different lineages can cocirculate at the same time in nature (27, 28), that genetically close viruses from the same lineage can circulate in a given environment at very different times in an apparent reemergence of ancient viruses (29), and that vaccination policies restrict diversity not only by stopping the spread of the virus in the field but in some cases by accidental reintroduction of the vaccine viruses themselves (16, 19, 29). In fact, in places in which FMD is enzootic, like South America, viruses show a much higher rate of accepted mutations and of variability, giving rise to rather complicated networks, whereas in Europe, where spreading of FMD is under control, the recorded sequences exhibit much less diversity. The 1979–1982 outbreak in Spain, documented by the CS sequences, and, in general, any outbreak of C1-type viruses that occurred in Europe since 1960 (34) consists of a reemergence of C1Obb-type genes (from 1960) and displays a tree-like evolution of little diversity. The same holds for the {OTh/81, OWupp/82} group, for the A5 viruses from 1983–1986 (revealing a reemergence of A5Ww-type genes observed in 1951), and even for the three O1-subserotype viruses from 1958, 1968, and 1982.

In summary, our diagrams document various forms of viral evolution and suggest the existence of quasispecies-like reservoirs of viruses, which—according to ref. 1—can account for the observed forms of evolution.

In addition, we hope to have demonstrated the suitability and the usefulness of split decomposition: a surprisingly high amount of the raw distance data can be processed this way. The resulting diagrams allow us to detect different evolutionary dynamics in given sequence families. We can distinguish different mutation rates and we can distinguish tree-like relations from less tree-like ones, both present in the same data.

We therefore expect our method to be useful as well to explain conflicting results in other studies [concerning, for example, the evolutionary position of *Prochloron* and *Prochlorothrix* (30) or of Archaebacteria (31, 32)] and for recognizing similarities resulting from convergence and, hence, in the long run for unraveling the functional role of nucleotide replacement. It also can be very useful in the study of the evolution of viroids, in which a kind of modular evolution, involving extensive recombination events of segments from different viroids, seems to occur (33).

We thank Drs. K. Strohmaier, E. Domingo, and F. Sobrino for many valuable comments. J.D. was supported by Sectorial grants in Science and Technology, Biotechnology Program, from the Commission of the European Communities. A.D. was supported by the Sonderforschungsbereich 7³ Diskrete Strukturen in der Mathematik, and A.v.H. was supported by the Forschungsschwerpunkt Mathematisierung.

1. Domingo, E. & Holland, J. J. (1988) in *RNA Genetics*, eds. Domingo, E., Holland, J. J. & Ahlquist, P. (CRC, Boca Raton, FL), Vol. 3, pp. 3–36.
2. Steinhauer, D. A. & Holland, J. J. (1987) *Rev. Microbiol.* **41**, 409–433.
3. Eigen, M., McCaskill, J. & Schuster, P. (1988) *J. Phys. Chem.* **92**, 6881–6891.
4. Eigen, M., McCaskill, J. & Schuster, P. (1989) in *Advanced Chemistry and Physics*, eds. Prigogine, I. & Rice, S. A. (Wiley, New York), Vol. 75, pp. 149–263.
5. Dress, A., von Haeseler, A. & Krüger, M. (1986) *Stud. Klassifikation* **17**, 229–305.
6. Fitch, W. M. & Margoliash, M. (1968) *Science* **155**, 279–284.
7. Hendy, M. D. & Penny, D. (1982) *Math. Biosci.* **59**, 277–290.
8. Penny, D. & Hendy, M. D. (1987) *Comput. Appl. Biosci.* **3**, 183–187.
9. Waterman, M. S., Smith, T. F., Singh, M. & Beyer, W. A. (1977) *J. Theor. Biol.* **64**, 199–213.
10. Bandelt, H.-J. & Dress, A. W. M. (1992) *Adv. Math.* **92**, 47–105.
11. Bachrach, H. L. (1968) *Annu. Rev. Microbiol.* **22**, 201–244.
12. Dopazo, J., Sobrino, F., Palma, E. L., Domingo, E. & Moya, A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6811–6815.
13. Domingo, E., Martinez-Salas, E., Sobrino, F., de la Torre, J. C., Portela, A., Ortin, J., Lopez-Galindez, C., Perez-Brena, P., Villaneuva, N., Najera, R., VandePol, S., Steinhauer, D., DePolo, N. & Holland, J. J. (1985) *Gene* **40**, 1–8.
14. Mateu, M. G., Martinez, M. A., Rocha, E., Andreu, D., Parejo, J., Giralt, E., Sobrino, F. & Domingo, E. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5883–5887.
15. Buneman, P. (1971) in *Mathematics in the Archeological and Historical Sciences*, eds. Hodson, F. R., Kendall, D. G. & Taute, P. (Edinburgh Univ. Press, Edinburgh), pp. 387–395.
16. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York), pp. 1–512.
17. Carrillo, C., Dopazo, J., Moya, A., Gonzalez, M., Martinez, M. A., Saiz, J. C. & Sobrino, F. (1989) *Virus Res.* **15**, 45–56.
18. Beck, E., Feil, G. & Strohmaier, K. (1983) *EMBO J.* **2**, 555–559.
19. Sobrino, F., Palma, E. L., Beck, E., Davila, M., de la Torre, J. C., Negro, P., Villaneuva, N., Ortin, J. & Domingo, E. (1986) *Gene* **50**, 149–159.
20. Beck, E. & Strohmaier, K. (1987) *J. Virol.* **61**, 1621–1629.
21. Bandelt, H.-J. & Dress, A. W. M. (1987) *Adv. Appl. Math.* **7**, 309–343.
22. Buonaugurio, D. A., Nakada, S., Parvin, J. D., Krystal, M., Palese, P. & Fitch, W. M. (1986) *Science* **232**, 980–982.
23. Buonaugurio, D. A., Nakada, S., Desselberger, U., Krystal, M. & Palese, P. (1985) *Virology* **146**, 221–232.
24. Li, W. H., Tanimura, M. & Sharp, P. M. (1988) *Mol. Biol. Evol.* **5**, 313–330.
25. Meyerhans, A., Cheynier, R., Albant, J., Seth, M., Kwok, S., Sninsky, J., Morfeld-Manson, L., Asjö, B. & Wain-Hobson, S. (1989) *Cell* **58**, 901–910.
26. Nichol, S. T., Rowe, J. E. & Fitch, W. M. (1989) *Virology* **168**, 281–291.
27. Rico-Hesse, R., Pallansch, M. A., Nottay, B. K. & Kew, O. M. (1987) *Virology* **160**, 311–322.
28. Pereira, H. G. (1981) in *Virus Diseases of Food Animals*, ed. Gibbs, E. P. G. (Academic, New York), Vol. 2, pp. 333–363.
29. Piccone, M. E., Kaplan, G., Giavedoni, L., Domingo, E. & Palma, E. L. (1988) *J. Virol.* **62**, 1469–1473.
30. Penny, D. (1989) *Nature (London)* **337**, 304–305.
31. Lake, J. A. (1988) *Nature (London)* **331**, 184–186.
32. Woese, C. (1987) *Microbiol. Rev.* **51**, 221–271.
33. Elena, S. F., Dopazo, J., Flores, R., Diener, T. O. & Moya, A. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5631–5634.
34. Martínez, M. A., Dopazo, J., Hernández, J., Mateu, M. G., Sobrino, F., Domingo, E. & Knowles, N. J. (1992) *J. Virol.* **66**, 3557–3565.