ELSEVIER

Biophysical Chemistry 66 (1997) 111–131

Biophysical
Chemistry

Graphs in Sequence Spaces: a Review of Statistical Geometry

Kay Nieselt-Struwe

Institut f. Zoologie, Univ. München, D-80021 München, Germany

Received 24 April 1997; accepted 24 April 1997

Abstract

Statistical geometry is a method of comparative sequence analysis of genes. Based on the concept of the sequence space of nucleic acids it computes the geometries of sequence sets, mainly quartets, by combining both the vertical and horizontal information content of the sequences. The geometries can be used to deduce, for example, the degree of tree-likeness of the data set without any *a priori* assumption of an evolution model. Furthermore, statistical geometry allows to detect varying positional substitution rates in sequences. Applications of the method to tRNA sequences have provided an assessment for the age of the genetic code. Furthermore, applications of statistical geometry to homeoboxes as well as different virus families have helped to assign reliable kinship relationships. In addition, a lower bound for the age of the common ancestor of the human and simian immunodeficieny viruses has been established.

Keywords: Statistical geometry in distance space; statistical geometry in sequence space; phylogenetic topologies; virus evolution; dating punctuation events

1. Introduction

This article reviews the last 15 years during which Eigen and co-workers have proposed and developed various comparative sequence analysis methods. Eigen's central idea to use bits of the positional information within a sequence alignment in order to study genetic evolutionary dynamics is most explicitly developed in the method of statistical geometry. This method was proposed as a tool complementary to the conventional sequence analysis procedures, in particular to the tree reconstruction methods. Since its first formal publication [1] applications to many biological data, especially viral data, have given useful insight into evolutionary dynamics. However, as with other methods that are complementary to the existing tree reconstruction methods, statistical geometry has since lived in the shadow of the latter ones. Presumably there are two main reasons for this: (i) due to the rather formal nature of the method, the results of an analysis with statistical geometry are often not easy to interpret, and (ii) because it does not produce a tree, it does not fulfil the expectation of most biologists who analyse sequence data. This article therefore also intends to make this elegant method more accessible to biologists.

One may classify the comparative sequence methods into distance space methods and sequence space methods. Distance space methods analyse sequences through their distance to other sequences based on a given metric. In sequence space the individual positions of a sequence alignment are analysed separately. Here is a list of the main methods that Eigen and co-workers have developed [2,3,1,4]:

PII S0301-4622(97) 00064-1

(1) Distance space methods
  (a) Pairwise distance statistics
  (b) Statistical geometry in distance space
(2) Sequence space methods
  (a) Consensus-deviation plots
  (b) Statistical geometry in sequence space
  (c) Monte Carlo plus statistical geometry

Viruses have proven to be among those systems that reflect clearly many of the characteristic laws of molecular evolution. Their fast mutation rate allows real-time observation of their evolution. Because of their fast replication and mutation rate viruses do not form families in the traditional population biological sense, but so-called quasispecies [5], a disperse distribution of similar sequences centered around the wild-type. This implies also that sequence relationships can often not be represented by a phylogenetic tree. Though there are methods to correct for the occurrence of parallel and back mutations the accumulation of noise cannot be compensated.
However, not all viruses behave in the same way. Quite the opposite has actually been observed. Despite the finding that most viruses operate near their error threshold, the patterns of mutation fixation can be quite diverse. In the following sections we will give some examples of this variability. We will see especially that the application of the method of statistical geometry can greatly enhance our knowledge of the evolutionary dynamics of RNA-viruses.

The article is organized as follows: After introducing the concept of the sequence space on which all comparative sequence analysis methods are based, the distance space methods that Eigen and co-workers have developed are reviewed. Then the sequence space methods will be summarized. Each method is first introduced rather formally, followed by many examples of applications to biological data. An overview of recent extensions and further developments of the method of statistical geometry will close this article.

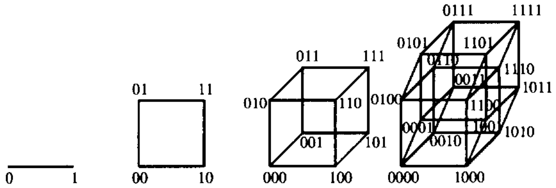

Fig. 1. The iterative built-up of the sequence space exemplified for binary sequences. Starting from one point each additional position (i.e., increase of the sequence length $\nu \mapsto \nu + 1$) leads to the doubling of all 2^{ν} points and the addition of further 2^{ν} edges that connect direct neighbors.

2. The concept of the sequence space

The idea of a sequence space was first introduced by Wright [6] and later used in information and coding theory by Hamming [7], [8]. The purpose is to represent all possible binary sequences of fixed length by a point space, such that each sequence is represented by a node, and an edge is drawn between any two nodes whose sequences differ by one bit. The number of different bits between any two sequences is also called the Hamming distance. In this point space the dimension corresponds to the number of positions in the sequences. For the construction of the ν-dimensional sequence space the following iterative procedure is intuitive (see figure 1, where this iterative procedure is shown for dimension 1 to 4): starting from a point (dimension 0), each consecutive dimension requires a doubling of the former diagram and subsequent connection of all new nearest neighbors by an edge.

The concept of the sequence space can of course formally also be applied to sequences such as proteins [9] with more than two or four symbols. Let κ be the number of symbols in the underlying alphabet.

There are some striking features about the sequence space, which elucidates its importance for the study of evolutionary processes:

- The cardinality of the sequence space and therefore the number of points is equal to κ^{ν}, which is enormous.
- The maximal distance between any two points in the space never exceeds the dimension ν of the space.
- There are $C(k) = \binom{\nu}{k}(\kappa - 1)^k$ different points with Hamming distance k from a given point.
- Between any two points separated by Hamming distance k there are $k!$ shortest paths, thus leading to tremendous connectivity.

In the case of the nucleotide alphabet, i.e. when $\kappa = 4$, the sequence space is constructed via a two-

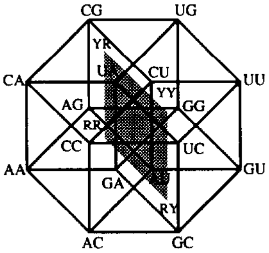

Fig. 2. The sequence space of nucleic acid sequences of length 2. R denotes the purine bases (A or G), Y the pyrimidine bases (C or U).

fold binary assignment process [10]. Each position is first assigned its base class, purine or pyrimidine, thereby a ν-dimensional hypercube is built, and then the nucleotide within each base class is assigned. Thus a second ν-dimensional hypercube at the point within the first ν-dimensional hypercube is generated. The resulting sequence space then has $4^{\nu} = 2^{2\nu}$ points and binary dimension 2ν. In figure 2 this construction process is exemplified for sequence length $\nu = 2$.

3. Comparative sequence analysis methods in distance space

The basis of comparative sequence analysis is the notion of distance. The prerequisite for the computation of the distance between two sequences is the alignment of the sequences. An alignment defines the evolutionary relationship of two homologous sequences and involves the identification of the locations of deletions or insertions that might have occurred in either of the two sequences. Since the problem of computing an optimal alignment of two or more sequences deserves an article by itself the reader is referred to the appropriate literature (see for example [11]). In the following it is assumed that an alignment is possible and given.

3.1. Pairwise distance statistics

Several evolutionary distance measures have been proposed for the comparative analysis of prealigned sequences. The simplest, the Hamming distance counts the number of positions at which the two sequences are occupied by different symbols. For two random sequences the expected Hamming distance is equal to 0.5ν for binary sequences, 0.75ν for quaternary sequences and for κ-letter alphabets in general it is $\frac{\kappa-1}{\kappa}\nu$.

The basis of the following procedure suggested by Eigen and co-workers [2] is the diffusion process which is modeled as follows: starting with n identical quaternary sequences of equal and fixed length ν, the sequences subsequently evolve independently and in parallel. We assume that the sequences may reproduce erroneously with a uniform substitution rate per site. Based on this model the following three distance values are considered:

(a) the average of the Hamming distances d_{i0} between individual sequences i and initial sequence 0,

(b) the average of the individual pair Hamming distances d_{ij},

(c) the Hamming distance d_{c0} between the consensus sequence c and initial sequence 0.

To establish a consensus sequence of the set one records for each position the symbol that appears most frequently. Of course, the resulting sequence is not necessarily unique in highly diverse populations. However, if the number of sequences in the alignment is odd and the alphabet binary then the consensus sequence is always unique.

Let $\overline{d_{i0}}(t)$ be the expected value for the mean Hamming distance at time t between any sequence and the ancestor sequence. Let $\overline{d_{ij}}(t)$ be the expected value for the mean Hamming distance at time t between any two sequences. Furthermore let $d_{c0}(t)$ be the expected value for the Hamming distance of the consensus and the initial sequence. Then for quaternary sequences of length ν the analytical expressions for these values are given by the following equations (P. Richter, personal communication; for a proof see also [12]):

$$\overline{d_{i0}}(t) = \frac{3}{4}\nu\left(1 - \exp\left(\frac{-4t}{3\nu}\right)\right) \qquad (1)$$

$$\overline{d_{ij}}(t) = \frac{3}{4}\nu\left(1 - \exp\left(\frac{-8t}{3\nu}\right)\right) \qquad (2)$$

$$d_{c0}(t) = \frac{\nu}{2}\operatorname{erfc}\left(\sqrt{\frac{n/6}{(1+3e^{\frac{-4t}{3\nu}}(1-e^{\frac{-4t}{3\nu}}}} \times \left(3e^{\frac{-4t}{3\nu}} - \sqrt{\frac{3}{2n}}\right)\right) \qquad (3)$$

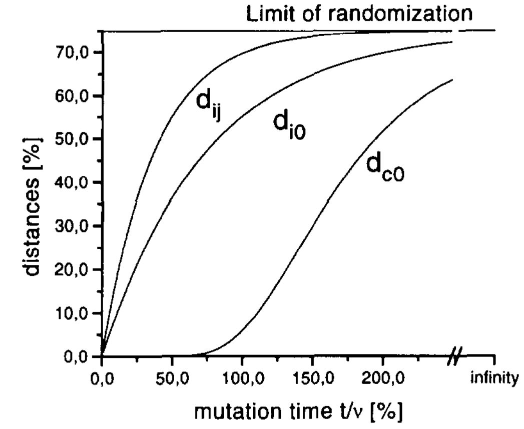

Fig. 3. Evolutionary divergence measured by the average distance of individual from initial sequence $\overline{d_{i0}}$, average pair distance $\overline{d_{ij}}$ and distance between consensus and initial sequence d_{c0} of sequences evolving according to the diffusion process. The curves represent the analytical formulae from equations (1)-(3).

where $\mathrm{erfc}(x)$ denotes the complementary error function defined as $1 - \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt$.

From the formula for the expected distance between consensus and initial sequence we can deduce that the two sequences coincide for quite a long time period before the consensus sequence starts to become more divergent (see figure 3). Since an experimental determination of the initial sequence will hardly be ever possible it is reasonable to compute the average distance of the individual sequences from the consensus sequence even if mutation times are large.

Vice versa, from a given mean Hamming distance $\overline{d_{ij}}$ we can compute the expected value of how many mutations have occurred:

$$t(\overline{d_{ij}}) = -\frac{3}{8}\mathrm{v}\ln\left(1 - \frac{3\overline{d_{ij}}}{4\mathrm{v}}\right) \qquad (4)$$

Similarly from a given mean Hamming distance $\overline{d_{i0}}$ we can deduce t:

$$t(\overline{d_{i0}}) = -\frac{3}{4}\mathrm{v}\ln\left(1 - \frac{3\overline{d_{ij}}}{4\mathrm{v}}\right) \qquad (5)$$

On the basis of the analytical formulae experimental sequence data can be analysed with respect to their relative noise level. If the values $t(\overline{d_{ij}})$ and $t(\overline{d_{i0}})$ are equal then there is good reason to believe that positions evolve uniformly according to the diffusion model.

Analysis of various tRNA sequences from different species according to this method showed an excellent correspondence between the mutation times t obtained from the experimental values $\overline{d_{ij}}$ and $\overline{d_{ic}}$ [2]. However, these two values could only be brought into conjunction after removing the constant positions from the alignment. It also revealed that the divergence of mitochondrial tRNA sequences has proceeded much further than that of eubacterial or eukaryotic species.

It should be noted that the distance analysis of DNA sequences based on the simple Hamming metric can be misleading if divergence has proceeded considerably. Then the chances of multiple substitutions at individual sites increases to the extent that the observed number of differences is likely to be smaller than the actual number of substitutions. Several models correcting for multiple substitutions have been proposed. Here we mention only the simplest one:

Let d be equal to the number of substitutions per site since the time of divergence between two sequences A and B.
The Jukes-Cantor distance [13] is given by

$$d_{\mathrm{JC}}(A,B) = -\frac{3}{4}\ln(1 - \frac{4}{3}p) \qquad (6)$$

where p is equal to the relative Hamming distance between the two sequences.

The Kimura 2-parameter distance [14] is yet a further generalization of the Jukes-Cantor model. It takes into account any difference in the rate of transitions and transversions. Let s and v be the proportions of transitional and transversional differences between the two sequences, respectively. Then

$$d_{\mathrm{K2P}}(A,B) = \frac{1}{2}\ln(a) + \frac{1}{4}\ln(b) \qquad (7)$$

where $a = 1/(1-2s-v)$ and $b = 1/(1-2v)$.

In the case of equal transition and transversion probabilities $v = 2s$ and the Kimura 2-parameter formula reduces to the Jukes-Cantor formula.

3.2. Statistical geometry in distance space

The basic idea of the method of statistical geometry is to provide a quantitative method for analysing the topology of branching. There are several reasons why sequence data might not have been generated by a tree-like process or why the reconstruction of the correct tree is difficult although the data have been generated by a tree-like process. Different tRNA sequences from a single organism are assumed to have a common ancestor of the same age as the genetic code. They are expected to have separated all more or less at the same time and evolved consequently independently and in parallel. Thus, their history is better represented by a bundle rather than by a binary tree. Non-tree relationships can also arise because of recombination, horizontal gene transfer, hybridisation or other reasons. Tree-like models can be blurred because of different rates of mutations along individual branches or within individual sequences.

Statistical geometry in distance space is based on the analysis of quartets of sequences. Two sequences define a unique pair distance, where the distance function is assumed to be arbitrary at first. Three sequences defining three pairwise distances can be expressed unambiguously in a distance diagram. However, in order to decide whether or not the three sequences share a precursor a fourth sequence needs to be added. One may suspect that the question whether a set of n sequences is tree-like or not can only be answered by looking at groups of more than just four sequences. Fortunately, it can be proved that it is sufficient to analyse quartet correlations in order to test for the overall tree topology, provided all $\binom{n}{4}$ quartets are taken into consideration (for the proof see [15], [16], [17]).

Consider four sequences A, B, C and D. Let δ be an arbitrary distance function. Then the four sequences define six pairwise distances $\delta(A,B)$, $\delta(A,C)$, $\delta(A,D)$, $\delta(B,C)$, $\delta(B,D)$ and $\delta(C,D)$, and three distance sums $\delta(A,B)+\delta(C,D)$, $\delta(A,C)+\delta(B,D)$, $\delta(A,D)+\delta(B,C)$. We order these three distance sums by magnitude and label them as L, M and S, with $L \geq M \geq S$.

The six distances can always be matched to a diagram with six segments (for a proof see [18]) resulting as solutions from the following equation system:

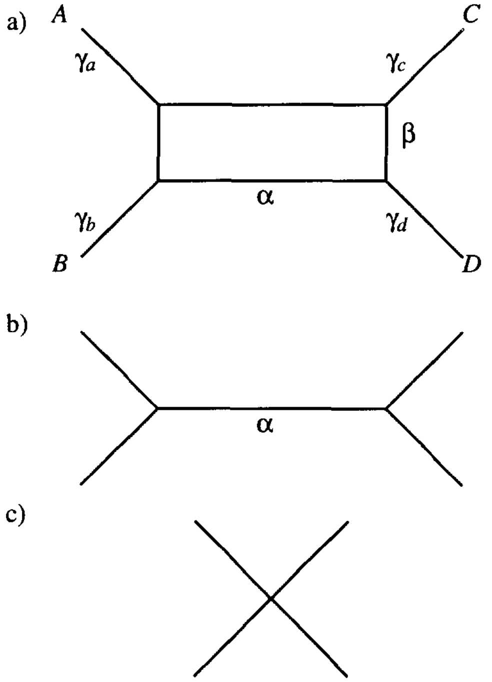

Fig. 4. Distance space graphs of 4 sequences. a) the general graph, b) ideal dendrogram ($\beta = 0$), c) ideal bundle graph ($\alpha = \beta = 0$).

$$\begin{aligned}
\gamma_A &= \frac{1}{2}(\delta(A,B)+\delta(A,C)+\delta(A,D)-L)\\
\gamma_B &= \frac{1}{2}(\delta(A,B)+\delta(B,C)+\delta(B,D)-L)\\
\gamma_C &= \frac{1}{2}(\delta(A,C)+\delta(B,C)+\delta(C,D)-L)\\
\gamma_D &= \frac{1}{2}(\delta(A,D)+\delta(B,D)+\delta(C,D)-L)\\
\alpha &= \frac{1}{2}(L-S)\\
\beta &= \frac{1}{2}(L-M)
\end{aligned} \tag{8}$$

The representative general diagram is shown in figure 4a. The distance segments γ_A, γ_B, γ_C, γ_D are the lengths of the four protrusions and α and β are the lengths of the inner rectangle.

We can distinguish the following cases: (1) all three distance sums are equal, then $L = M = S$ and therefore $\alpha = \beta = 0$, (2) two distance sums are equal and larger than the third, then $L = M > S$ and thus

$\beta = 0, \alpha > 0$ and (3) all three distance sums differ, then $L > M > S$ and therefore $\beta > 0, \alpha > 0$. In case (1) the resulting diagram is a star graph representing an ideal bundle (fig. 4c) and in case (2) the resulting diagram is a perfect dendrogram representing an ideal tree (fig. 4b). A positive value β and/or the ratio β/α is therefore a measure of "deviation from ideal tree-likeness". For random sequences this ratio is equal to 0.5.

For a set of n sequences the method becomes statistical by looking at all possible $\binom{n}{4}$ quartets and constructing the average distance diagram. As mentioned above one can prove that the set is an ideal tree if each of the quartets are ideally tree-like, that is if all $\binom{n}{4}$ values β are zero and all values α are positive. Hence, the averages $\overline{\beta}$ and $\overline{\alpha}$ and/or the ratio $\overline{\beta}/\overline{\alpha}$ are qualitative measures how much the sequence set deviates from an ideal tree.

For a set of binary sequences evolving according to the diffusion process as modeled in the previous section the values of the mean protrusion lengths as well as the rectangle lengths tend with increasing mutation divergence toward a saturation [19]. The value $\overline{\beta}$ approaches quickly its saturation point, thus showing how insensitively it reacts to randomization.

The mean distance space graphs of two different virus families are shown in figure 5. In figure 5a the NS gene of 15 DNA-sequences of Influenza A isolated between 1933 and 1985 was analysed. In figure 5b the average distance diagram of a set of Polio DNA-sequences is shown. The data set was taken from [20] and comprised 57 sequences of Polio type 1 with 150 bases that were isolated in 5 different continents during a span of 31 years.

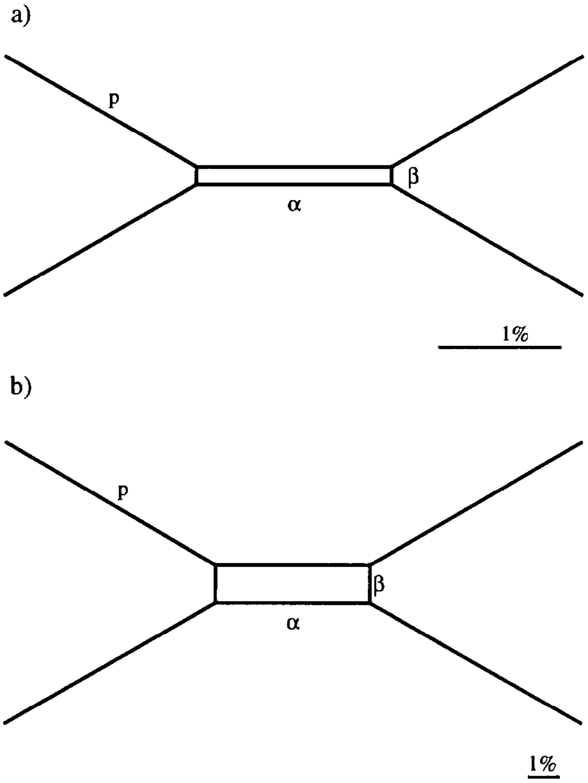

Fig. 5. The mean distance space graphs of (a) 15 Influenza-A DNA-sequences and (b) 57 Polio DNA-sequences.

3.3. *Testing phylogenetic trees*

A modification of the statistical geometry in distance space was proposed by Rodrigo and Dopazo [21]. The authors applied the method to test each individual inner branch of a given computed phylogenetic tree. In this case not all $\binom{n}{4}$ quartets that can be chosen from a set of n sequences are analysed, but only those that arise from the clustering defined by the respective branch (figure 6). The six pairwise distances of each quartet given by the branching order are computed, and the parameters β and α are obtained from these data. The ratio $\overline{\beta}/\overline{\alpha}$ of the mean values then indicates the amount of "noise" in the data, and could therefore be used to quantify the reliability of that branch. The authors coined the term branch noise analysis (BNA) to characterize this method.

An exhaustive evolutionary analysis of the picornavirus family using the amino acid sequences of several proteins was carried out by Rodrigo and Dopazo in order to define unambiguously the different genera within this family. Several methods besides the BNA method were used to test the reliability of the computed phylogeny. In their study the authors showed that there is a clear relationship between unreliable branching points in the picornavirus family (as shown independently by the bootstrapping procedure [22]) and large degree of evolutionary noise (as indicated by $\overline{\beta}/\overline{\alpha}$ ratios close to 0.5).

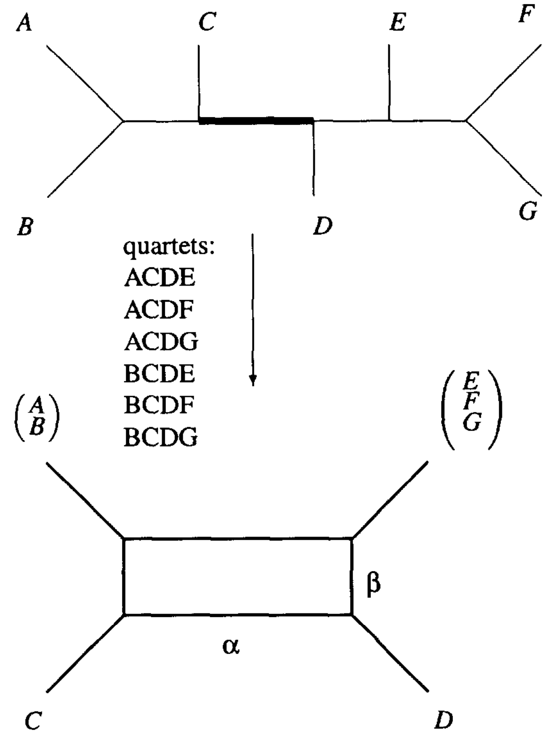

Fig. 6. Testing the reliability of a specified branch of a phylogenetic tree. The average tree-likeness parameter $\bar{\beta}/\bar{\alpha}$ of the statistical geometry in distance space of all quartets arising from the branching order as shown is used.

4. Comparative sequence analysis methods in sequence space

4.1. Consensus-deviation method

The following procedure enables the detection of variable positions. Given an alignment S of n sequences of length ν we first determine the consensus sequence, denoted as M_S. We assume that the consensus sequence M_S is unique. For every position j in the alignment let δ_j be equal to the number of sequences that differ from the consensus sequence at position j. The frequency distribution $f(\delta_j)$ of the positional distances δ_j then provides information about the degree of divergence in the given data set. Thus a distribution of binary sequences with a peak at $\delta = 0.5n$ signals complete randomization. In figure 7 histograms for $f(\delta)$ obtained by simulating a diffusion process of binary sequences are shown.

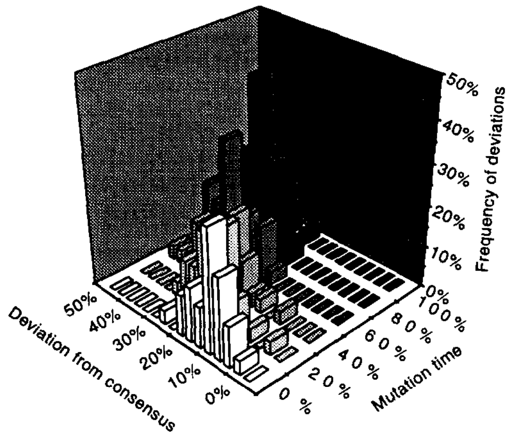

Fig. 7. Plot of the frequency of deviations from the consensus. A diffusion-like divergence process with uniform probabilities of substitution was simulated using $n = 30$ sequences of length $\nu = 1000$.

By applying this procedure to various sets of tRNA sequences it was possible to distinguish constant from variable positions in the tRNA alignment ([2], [3], [23], [24]). It showed that most of the constant positions are situated in the loops and most of the variable positions can be found in the stem regions. The constant positions do not necessarily represent ancestral information but rather functional constraints. The variable positions, which could further be distinguished as moderately and highly variable, encode functions that require discrimination of individual tRNAs.

4.2. Statistical geometry in sequence space

The formal definition of the method of statistical geometry in sequence space was published by Eigen and co-workers in 1988 [1] (reviewed in [19]). This method was proposed as an extension of statistical geometry in distance space. Statistical geometry in sequence space is a model-free procedure for evaluating the underlying phylogenetic relationships among a sequence set based on the position-wise entries of the sequences. It was first introduced using the unweighted Hamming metric [1], and was later extended to the use of arbitrary metrics ([25], Nieselt-Struwe *et al.*, submitted).

4.3. Statistical geometry with the Hamming metric

Let us illustrate the principle of the statistical geometry in sequence space for the case of four binary sequences and the Hamming metric. Let A, B, C, D be four sequences of length ν from a binary alphabet $\mathcal{A}$. We denote the letters of the alphabet by 0 and 1. We imagine the four sequences to be written in the form of a matrix, where the rows are the sequences, and the columns refer to the positions in the sequences. We first want to determine all the possible combinations of "equal/non-equal" entries of four binary sequences when regarding only a single column. We can distinguish eight such combinations if the order of the sequences is kept (figure 8): in the first category the four sequences have equal entries, then there are four categories where one of the four sequences differs from the other three and finally there are three categories in which there are two pairs of equal entries. For a given sequence alignment we sum up the number of positions in each of the eight categories. For these sums let us introduce the following notation: we count the number of positions at which the entries of the four sequences are equal. We denote this number by $g(ABCD)$. Similarly, we count the number of positions at which the entries of sequence A differ from the three sequences B, C and D, and denote this number by $g(A|BCD)$. In the same way $g(AB|CD)$ counts the number of positions where A and B as well as C and D have distinct pairwise equal entries.

We call the parameters $g(.|.)$ the sequence space parameters of the quartet. Note that the sequence length ν is equal to the sum of all eight sequence space parameters $g(.|.)$.

The remarkable feature of these eight parameters is that neither permutations of the positions of the alignment nor permutations of the elements of the alphabet change them (for a proof see [12]). Thus the sequence space parameters are so-called invariants. As an example let us look at the following four sequences:

$$\begin{aligned} A &= 0110 \\ B &= 0101 \\ C &= 0011 \\ D &= 0000 \end{aligned}$$

	d_0	$\underbrace{\sum = d_1}$				$\underbrace{\sum = d_2}$		
	o	a	b	c	d	x	y	z
A	0	1	0	0	0	1	1	1
B	0	0	1	0	0	1	0	0
C	0	0	0	1	0	0	1	0
D	0	0	0	0	1	0	0	1

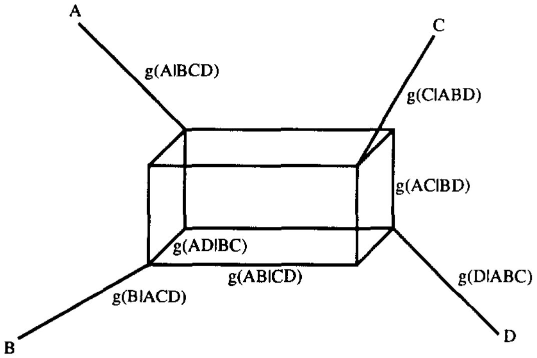

Fig. 8. The sequence space graph of quartets of binary sequences. There are eight position categories: in $o = (ABCD)$ all four sequences have equal entries, in $a = (A|BCD)$, $b = (B|ACD)$, $c = (C|ABD)$, $d = (D|ABC)$ one sequence differs from the other three, in $x = (AB|CD)$, $y = (AC|BD)$, $z = (AD|BC)$ there are two pairs of sequences with equal entries. Summing up the number of positions in the alignment in each of the eight position categories gives rise to the eight distance segments $g(.|.)$. The seven distance segments referring to the non-homogeneous positions can be represented by a graph as shown below, where the exterior vertices correspond to the four sequences and the lengths of the edges are equal to the respective distance segments.

Then

$$\begin{aligned} g(ABCD) &= 1 \\ g(A|BCD) = g(B|ACD) &= \\ g(C|ABD) = g(D|ABC) &= 0 \\ g(AB|CD) = g(AC|BD) = g(AD|BC) &= 1 \end{aligned}$$

Now we permute the columns of the alignment by $\pi = (2431)$ and exchange 0 by 1 within each column. This yields the four sequences

$$\begin{aligned} A' &= 1100 \\ B' &= 0110 \\ C' &= 0101 \\ D' &= 1111 \end{aligned}$$

and clearly the same sequence space parameters:

$$\begin{aligned} g(A'B'C'D') &= 1 \\ g(A'|B'C'D') = g(B'|A'C'D') &= \\ g(C'|A'B'D') = g(D'|A'B'C') &= 0 \\ g(A'B'|C'D') = g(A'C'|B'D') = g(A'D'|B'C') &= 1 \end{aligned}$$

The next step is to represent the sequence space parameters by a graph (figure 8). This graph is now a box, whose dimensions are equal to $g(AB|CD)$, $g(AC|BD)$ and $g(AD|BC)$, with four protrusions (referring to $g(A|BCD)$ etc.) connecting to the four exterior vertices, which represent the four sequences. As in the case of the distance space graph, the underlying topology can be deduced from the sequence space graph: while an ideal tree will show only one positive box dimension, the box dimensions of an ideal bundle are all zero, and the box dimensions of four random sequences, representing a netted topology, are all of similar magnitude.

The extension of the procedure to larger alphabets is straightforward. For a quaternary alphabet such as the nucleic acid alphabet in addition to the eight sequence space parameters seven further ones have to be added: there are six position combinations in which there is exactly one pair of equal entries (yielding $g(AB|C|D)$, $g(AC|B|D)$, $g(AD|B|D)$, $g(BC|A|D)$, $g(BD|A|C)$, $g(CD|A|B)$) and one category of all unequal entries (leading to $g(A|B|C|D)$). The representative graph in which the sides of the triangles comprise $5/6$ of the average of the six tetrahedral dimensions plus the value of $g(A|B|C|D)$ is shown in figure 9.

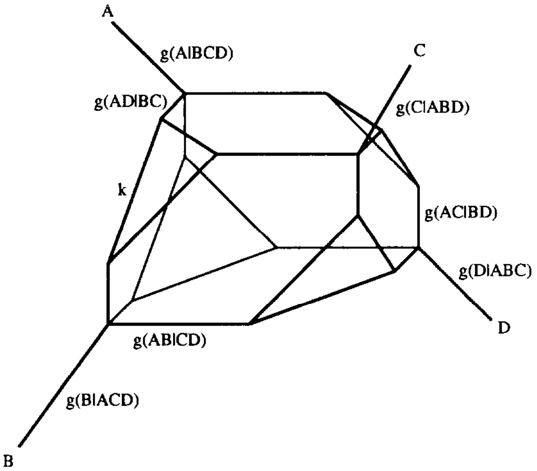

Fig. 9. The sequence space geometry of a quartet of quaternary sequences. In addition to the positions categories of binary sequences, there is one position category with six realizations in which there is exactly one pair of equal entries and one position category in which all entries differ from each other. k is then equal to 5/6 of the average of the tetrahedral dimensions plus the number of positions with all different entries.

How does the method extend to sets of $n > 4$ sequences? As in the case of the statistical geometry in distance space, for each quartet the sequence space parameters are computed. Since the order of the sequences can no longer be kept, the four protrusion parameters are ordered by magnitude, as well as the three box dimensions, denoted by $L(box)$, $M(box)$, $S(box)$, and (for quaternary alphabets) the six tetrahedral dimensions, denoted by $L(tetra)$, $M(tetra)$, $S(tetra)$. In addition we compute the sums of the sequence space parameter within each position class as follows:

$$d_0 = g(ABCD)$$

is equal to the number of homogeneous positions,

$$\begin{aligned} d_1 &= g(A|BCD) + g(B|ACD) + g(C|ABD) \\ &\quad + g(D|ABC) \end{aligned}$$

is equal to the number of positions for which one sequence differs from the other three,

$$d_2 = g(AB|CD) + g(AC|BD) + g(AD|BC)$$

is equal to the number of positions with two pairs of equal entries. In addition, for quaternary sequences

$$d_3 = g(AB|C|D) + \ldots + g(AD|B|C)$$

is equal to the number of positions with exactly one pair of equal entries, and finally

$$d_4 = g(A|B|C|D)$$

is equal to the number of positions with no pair of equal entries.

Then the averages of each of the sequence space parameters for all $\binom{n}{4}$ quartets are computed, leading to the statistical sequence space parameters. The corresponding graph is called the statistical geometry of the sequence set.

From this graph the underlying global topology of the sequence set can be deduced: a tree-like topology will show one large box dimension (and two large tetrahedral dimensions for quaternary sequences), while a bundle-like topology has inner box and/or tetrahedral dimensions of similar magnitude.

Formally we define the tree-likeness T and the bundle-likeness B of the sequence set by:

$$\mathsf{T} = \frac{\overline{L}(box) + \overline{L}(tetra)}{\overline{d_2} + \overline{d_3}} \tag{9}$$

$$\mathsf{B} = \frac{\overline{L}(box) + \overline{S}(box) + \overline{L}(tetra) + \overline{S}(tetra)}{2 \cdot \overline{M}(box) + \overline{M}(tetra)} \tag{10}$$

where $\overline{d_2}$, $\overline{d_3}$ denote the averages of the sum of the three box dimensions and the averages of the sum of the six tetrahedral dimensions respectively. Note that in binary sequence space the tetrahedral parameters are equal to zero.

For sequence sets that can be represented by ideal trees, $\mathsf{T} = 1$ and $\mathsf{B} = \infty$, whereas sequence sets generated by diffusion-like processes yield ideal cube-like graphs with $\mathsf{B} = 1$ and $\mathsf{T} = 1/3$.

As examples of model topologies a bundle-like divergence and two tree models were analysed. Figure 10 shows the average sequence space parameters $\overline{d_0}, \ldots, \overline{d_4}$ as functions of relative mutation distance obtained from analytical formulae [12] assuming parallel and independent divergence with uniform substitution rates.

The sequences of the two tree models were generated using the program Seq-Gen [26]. This program simulates the evolution of nucleotide sequences along a specified phylogeny given some model substitution process. The models and the corresponding quaternary statistical geometry graphs are shown in figure 11.

As biological examples, the geometries of the DNA sequences as well as deduced purine/pyrimidine sequences of the Influenza virus and the Polio virus as used above are shown in figure 12.

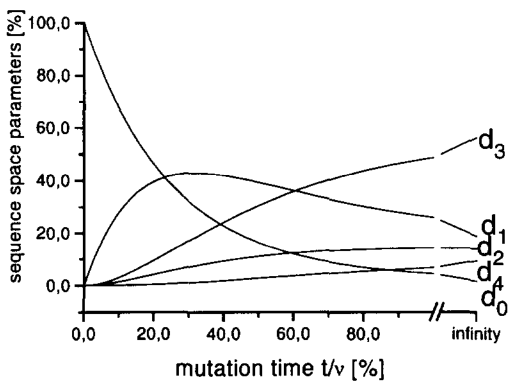

Fig. 10. The five statistical sequence space parameters $\overline{d_0}, \ldots, \overline{d_4}$ of quartets of quaternary sequences as a function of the relative mutation time t/v. A diffusion-like process with uniform substitution probabilities was assumed. The values $\overline{d_i}$ were calculated using the analytical formulae derived in [12].

The sequence space diagrams of the Influenza virus are almost ideally tree-like, thus confirming the time-resolved tree according to Buonagurio and co-workers [27]. Based on that tree the authors determined an average rate of fixation of 1.9 mutations per 1000 residues per year, a result also found by Eigen and Nieselt-Struwe [4] using the temporal increase of the box dimensions of adjacent quartet combinations (i.e. of all individual quartets that are separated by an edge in the time-resolved tree).

The divergence pattern of the Polio virus is entirely different from that of the Influenza virus. While the first and the second codon positions appear almost invariant, the third codon position is largely randomized. Nearly every mutation changing the amino acid sequence of this region would apparently be lethal for the virus.

Eigen and co-workers applied the method of statistical geometry to various tRNA sequences [24]. They showed that those tRNA sequences corresponding to a specified amino acid from different organisms form a tree-like topology, whereas all the tRNA sequences from a single organism form a bundle-like topology. The root of these bundles mark the time of fixation of the genetic code and/or the time of assignment of amino acids to their cognate tRNAs. From their

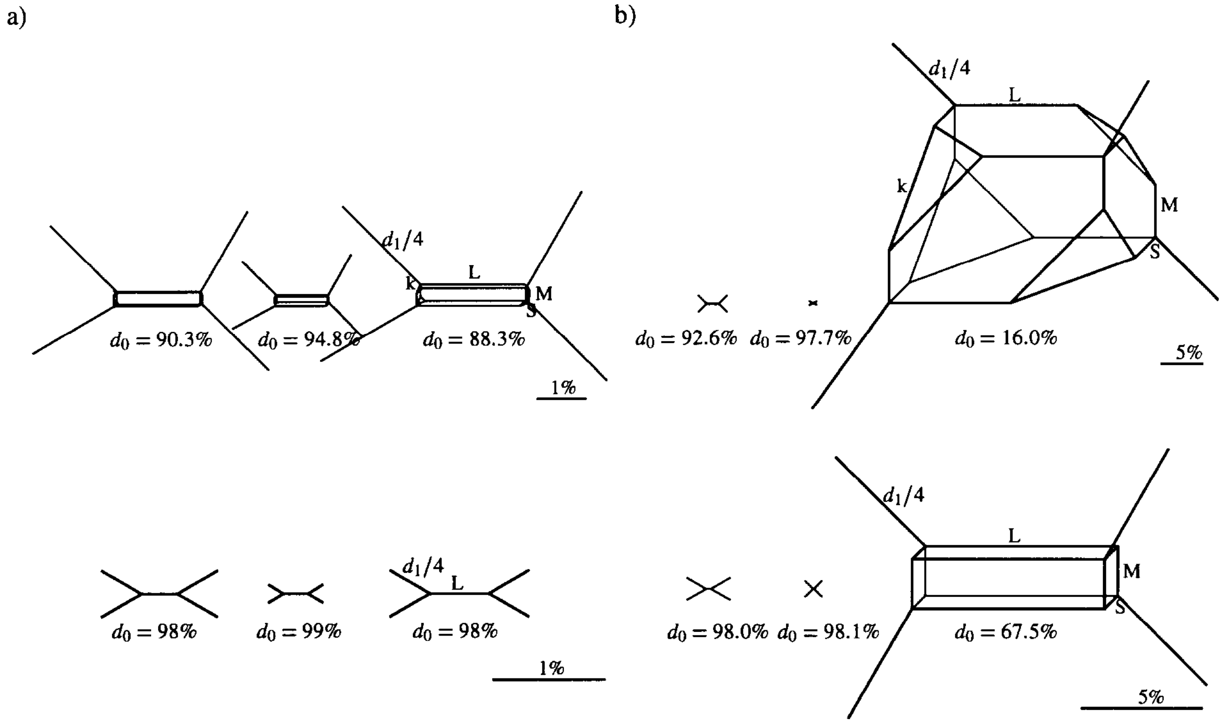

Fig. 12. Statistical geometries of the influenza A virus and the Polio-1 virus. The three graphs in each row refer to the statistical geometries of the first, second and third codon position of the gene considered. The upper row in (a) and in (b) shows the graphs of the DNA-sequences, the lower row represents the graphs of the sequences translated into the purine/pyrimidine sequence space. The parameter d_0 is equal to the relative mean number of homogeneous positions in each quartet in percent of the sequence length. In (a) a unit length referring to 1% distance (relative to sequence length) is shown, for (b) the scale of 5% distance is drawn. (a) Mean geometries of 15 sequences the NS gene of Influenza A. (b) Mean geometries of 57 sequences of a region of the VP1/2A gene of Polio-1.

studies the authors concluded that the genetic code is not older than 3.8 billion years and therefore younger than the age of our planet (being approximately 4.7 billion years).

The method of statistical geometry allows not only the computation of the DNA geometries or of their deduced purine/pyrimidine sequences, but also the direct visualization of the geometries based merely on transitional mutations. It also allows the calculation of the ratio of transitional changes versus transversional changes. The procedure is based on the construction of the nucleic acid sequence space as a two-fold binary hypercube. In order to compare the degree of transversional divergence with the degree of transitional divergence we first translate the quartet of DNA-sequences into the purine/pyrimidine sequences. For these sequences the sequence space parameters d_0, d_1 and d_2 are computed. Then the relative number of transversional changes is equal to

$$v = \frac{d_1 + d_2}{\nu} \tag{11}$$

where ν is equal to the sequence length.

Now we consider all the positions that are homogeneous in the RY-space, i.e., all positions which sum up to d_0. Any changes in this subspace on the DNA-basis refer exclusively to transitional changes and thus give rise to a binary sequence space graph. We denote the corresponding sequence space parameters by d_{00}, d_{01} and d_{02}, where d_{00} is equal to the number of homogeneous positions in the subspace of transitions, d_{01} is equal to the number of positions where one entry differs from the other three and d_{02} is equal to the number of positions with pairwise

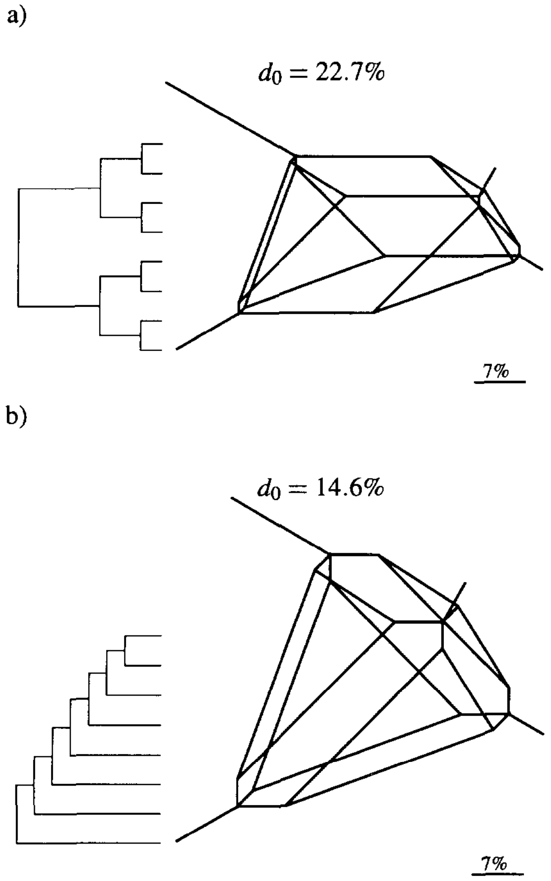

Fig. 11. The statistical geometries of two tree models as shown to the left. The corresponding DNA-sequences were generated using Seq-Gen, a Monte Carlo type simulation program [26]. Independent and uniform evolution with equal transition and tranversion probabilities was assumed. In both models the length of the sequences was equal to 1000.

equal transitional entries. The first index 0 indicates that the sequence space is d_0-dimensional. The relative number of transitional changes is consequently computed by

$$s = \frac{d_{01} + d_{02}}{d_0} \tag{12}$$

For sets of more than four sequences the average ratio of transitions to transversions is deduced from the average parameters s and v from each of the quartets considered.

In table 1 some examples of computed ratios of transitions versus transversions from virus sequences are shown.

Table 1
Ratio of transitional to transversional changes in Influenza A and Polio type 1 virus. The ratio was determined for each codon position of the genes considered.

Codon position	Influenza A	Polio 1
1	4,0	2,8
2	5,2	0,2
3	4,5	2,3

4.4. Statistical geometry with arbitrary metrics

So far we have defined the method of statistical geometry in sequence spaces with a Hamming metric. But the Hamming metric is not such a useful metric for the examination of amino acid sequences, for example. In this section we present an extension of statistical geometry to sequence spaces with arbitrary dissimilarity functions between the symbols of the alphabet. This generalized method allows, for example, the analysis of amino acid sequences with Dayhoff-like or other substitution matrices.

As we have shown in section 3 four sequences A, B, C, D with any dissimilarity function δ yield a canonical distance space graph (fig. 4). In order to apply this concept to arbitrary sequences we assume that a reasonable distance function δ between the symbols of the alphabet is given. In 1978 Dayhoff and co-workers [28] suggested such an exchange matrix for amino acids, which was later modified by Gonnet and co-workers [29]. Each position in the alignment of the four aligned sequences A, B, C, D each of length ν is analysed separately as follows: Assume that at position I the entries of A, B, C, D are w, x, y, z respectively. Then $\delta(A_I, B_I) = \delta(w, x)$, $\delta(A_I, C_I) = \delta(w, y)$ and so on. Thus we can compute six pairwise distances and the three distance sums. The three distance sums are ordered by magnitude and the six distance segments γ_A, γ_B etc. of the distance space diagram are deduced. On the other hand we can identify the distance space diagram with the binary sequence space diagram where the third box dimension has length zero in the following way:

At position I, $I \in \{1, \ldots, \nu\}$ let L_I be equal to the maximum of the three distance sums. Then we define

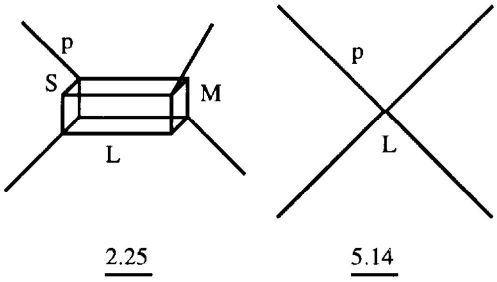

Fig. 13. Weighted statistical geometries of the Polio-1 virus. The same sequences as in figure 12 were used. In (a) the DNA-geometry was calculated using a dissimilarity function on the ATGC-alphabet weighing transversions 4 times higher than transitions. In (b) the weighted statistical geometry of the deduced amino acid sequences was computed using the Gonnet exchange matrix. For each graph a unit length referring to 5 times the mean distance of the dissimilarity values is shown.

$$
\begin{aligned}
g_I(AB|CD) &:= \frac{1}{2}(L_I - \delta(A_I,B_I) - \delta(C_I,D_I)) \\
g_I(AC|BD) &:= \frac{1}{2}(L_I - \delta(A_I,C_I) - \delta(B_I,D_I)) \\
g_I(AD|BC) &:= \frac{1}{2}(L_I - \delta(A_I,D_I) - \delta(B_I,C_I)) \\
g_I(A|BCD) &:= \gamma_A \\
g_I(B|ACD) &:= \gamma_B \\
g_I(C|ABD) &:= \gamma_C \\
g_I(D|ABC) &:= \gamma_D
\end{aligned}
$$

This is done for each position in the alignment and the respective distance segments are summed up, yielding $g(\cdot|\cdot) = \sum_I g_I(\cdot|\cdot)$.

The corresponding weighted graph for A,B,C,D with the four protrusion segments $g(A|BCD)$, $g(B|ACD)$, $g(C|ABD)$, $g(D|ABC)$, and the three box dimensions $g(AB|CD)$, $g(AC|BD)$, $g(AD|BC)$ is similar to the binary sequence space graph as shown in figure 8.

An extension to sets of more than 4 sequences is done equivalently to the unweighted case, as is the definition of the tree-likeness and bundle-likeness of the set.

As an example let us reexamine the Polio sequences used above. Since the ratio of transitional versus transversional mutations was computed to be approximately 2 for the third codon position, the weighted statistical geometry taking this ratio into account was computed. The corresponding weighted statistical geometry is shown on the left in figure 13. In addition, the statistical geometry of the deduced amino acid sequences using the Gonnet exchange matrix [29] was calculated (right graph in figure 13). The geometry is now ideally tree-like, the inner box dimension however is small in comparison to the protrusions.

4.5. *Statistical geometry and the reliability of phylogenetic trees*

The method of statistical geometry in sequence space cannot only be used for the *a priori* assessment of the topological structure of a sequence set, it is also a successful tool to test the hypothesis of a defined clustering. The latter can readily be applied to test proposed branching of a given phylogenetic tree, similar to the procedure described in section 3.3. In other words it is also an *a posteriori* method for testing the reliability of a reconstructed tree. The procedure is evident: The sequence set is partitioned into four subsets. Let these four subsets contain n_1, n_2, n_3 and n_4 sequences respectively. Then the sequence space parameters of the $n_1 \cdot n_2 \cdot n_3 \cdot n_4$ possible quartets are computed and from these the relevant averages are deduced. If one of the inner box dimensions and/or tetrahedral dimensions is much larger than the other ones, then this is an indication that two groups share a common ancestor.

In figure 14 the statistical geometries of five human and simian immunodeficiency virus (HIV/SIV) groups in the purine/pyrimidine sequence space and in the amino acid sequence space are shown. The hypothesis was tested that all HIV and SIV groups share a common ancestor from which the five groups evolved rapidly and independently, and therefore the sequences in the earliest node are best represented by a bundle-like topology. From the five groups $\binom{5}{4} = 5$ partitions into four clusters can be chosen, and for each of them the statistical geometry is shown. The gag gene was chosen, a gene encoding the core of the virus, and after removing all gap positions from the alignment the sequences were translated into the purine/pyrimidine sequences. In addition, the geometries of the deduced amino acid sequences were computed using the Gonnet exchange matrix. From the

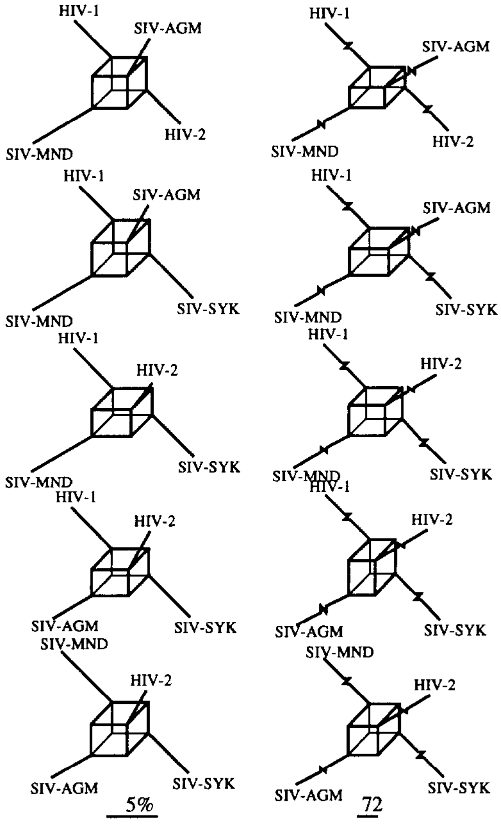

Fig. 14. Graphs of the statistical geometry analysis of 1254 positions of the *gag* gene of HIV-1/HIV-2/SIV-AGM/SIV-MND/SIV-SYK. The five groups were combined into the $\binom{5}{4} = 5$ possible 4-partitions. The five diagrams on the left represent the geometries of the RY-sequences using the Hamming metric, on the right the geometries of the amino acid sequences using the Gonnet exchange matrix are shown. For the left column a unit length referring to 5% (relative to the total sequence length) is shown. For the right column a unit length referring to 5 times the mean distance of the entries in the Gonnet matrix is shown.

geometries it is evident that indeed all five groups show no obvious branching order. Similar results have been obtained using the DNA-sequences as well as the sequences of other genes of the virus, such as the gene for the envelope protein.

By combining tree reconstruction methods with this adaptation of the method of statistical geometry, Schubert and co-workers [30] were able to reconstruct reliable phylogenetic trees of homeobox genes both from insects and vertebrates. The remarkable linear order of homeobox genes both in insects and vertebrates, as well as the close correlation of the genes' expression boundaries and their position on the chromosome, have challenged scientists to explain the evolutionary steps leading to this organization. From their studies Schubert and co-workers concluded that there was (i) a common ancestor for all vertebrate homeobox gene clusters, from which the clusters evolved by duplication, and (ii) a subdivision of the vertebrate and insect homeobox genes into three classes corresponding to the three main regions of the embryo. From this ancient three-gene cluster subsequent duplication events led to a cluster of at least five homeobox genes in the common ancestor of vertebrates and insects.

4.6. *Monte Carlo and statistical geometry*

A Monte Carlo type method of resampling was introduced by Archie [31] and Faith ([32]; see also [33]) and its application to statistical geometry was proposed by Eigen and co-workers [1]. It is used to test the hypothesis that there is no tree-likeness in the data and involves the successive permutation of the entries in the columns of the sequence alignment. This produces alignments of the same number and kinds of characters but no taxonomic structure. Note that this procedure leaves (i) the consensus sequence and (ii) the mean Hamming distance unchanged. One computes then the bundle-likeness parameter B as a function of the number of changes. If that statistic then shows a tendency to decay toward a value close to one, then there is some taxonomic structure in the data (though perhaps it might be just a pair of sibling species). If on the other hand the quantity fluctuates randomly no residual tree-likeness is present in the data.

Figure 15 demonstrates the application of this procedure to the Polio sequences as used in the previous sections. The third codon position of the purine/pyrimidine sequences were successively shuffled within each of the 50 positions, and the bundle-likeness parameter B as defined in equation (10) was plotted against the relative number of shuffling steps. The plot clearly shows a decay of the bundle-likeness parameter toward $B = 1.0$, which is indicative of a certain degree of residual tree-likeness in the data.

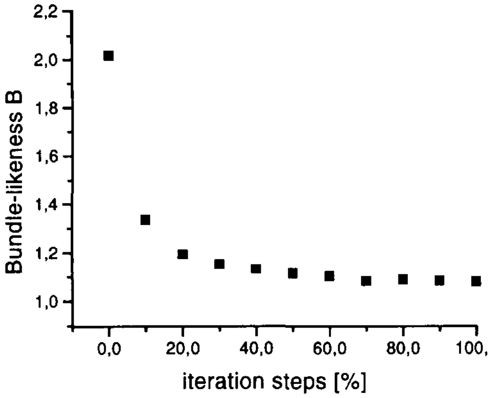

Fig. 15. Applying the Monte Carlo reshuffling procedure to the third codon position of the purine/pyrimidine sequences of Polio-1. The bundle-likeness parameter B was computed as a function of the shuffling steps.

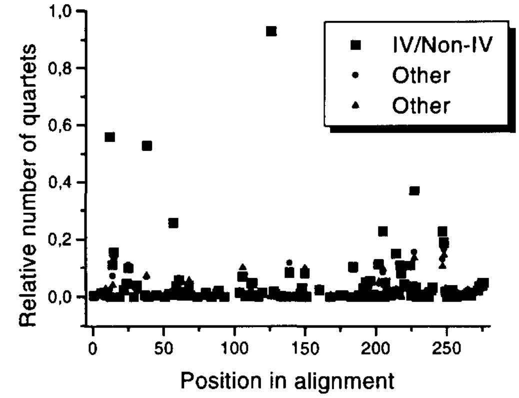

Fig. 16. Positional statistical geometry of 72 HIV-1 sequences from different risk groups. The support of the clustering of the sequences into two groups was tested, one consisting of the drug user-derived sequences (IV), the other one of the rest (Non-IV). At each position of the 260 positions in the alignment the relative number of quartets supporting the hypothesis is plotted.

4.7. *Positional analysis*

Another useful feature of statistical geometry is the capability to test individual positions with respect to their variability, tree-likeness or support of a given clustering. Rather than analysing the columns of each quartet, and summing up the number of positions in each of the position categories, each column of the whole alignment is investigated separately. This is equivalent to the length of the alignment being equal to one. At each position the averages of the sequence space parameters of all quartets to be considered are computed, from which the qualitative parameters such as tree-likeness or support for the clustering is deduced.

In figure 16 a data set of 72 human immunodeficiency virus type 1 sequences has been analysed using the positional method. This set consists of a large number of V3 sequences from different risk groups and different years. The V3 region contains an important neutralization epitope; it is one of the hypervariable regions in the external envelope of the virus. Extensive analyses of the data were carried out by Kuiken and co-workers by using the method of statistical geometry to assess the amount of tree-likeness in the sequences [34]. Phylogenetic analyses of the data showed only one obvious and epidemiologically meaningful cluster division, namely the distinction of drug user-derived sequences and the other ones (from homosexuals and hemophiliacs). The application of CART, a method designed explicitly to detect the positions that support a given clustering [35], showed that only one position is used for the discrimination between the two groups. The position plot in figure 16 not only supports this result, but also shows directly that no other clustering is supported in any other position.

5. The diffusion model

In this section we want to have a closer look at bundle topologies. Underlying such a bundle graph is the diffusion process. As we have demonstrated for some examples from different virus sequences, the bundle graph describes most appropriately the phylogenetic relationships governed by a diffusion process.

As above our model system consists of n binary sequences of equal and fixed length ν, which have separated from a common ancestor at time $t = 0$, and evolved independently and in parallel. We assume that the sequences may reproduce erroneously with a uniform positional substitution rate. This is a reasonable assumption, at least for non-coding sequences.

In the following we want to write down analytical expressions for the binary sequence space parameters $\overline{d_0}(t)$, $\overline{d_1}(t)$ and $\overline{d_2}(t)$ at time t. Again t is chosen such that on average one mutation per sequence per time unit appears. Then

$$\overline{d_0}(t) = \frac{1}{8}\left(1+6e^{\frac{-4t}{\nu}}+e^{\frac{-8t}{\nu}}\right) \tag{13}$$

$$\overline{d_1}(t) = \frac{1}{2}\left(1-e^{\frac{-8t}{\nu}}\right) \tag{14}$$

$$\overline{d_2}(t) = \frac{3}{8}\left(1-e^{\frac{-4t}{\nu}}\right)^2 \tag{15}$$

For $t > 0$ let

$$r_1(t) = \overline{d_1}(t)/\overline{d_0}(t)$$

and

$$r_2(t) = \overline{d_2}(t)/\overline{d_1}(t).$$

Given r_1 and/or r_2 we are interested in t:

$$t(r_1) = -\frac{\nu}{4}\ln\left(\frac{-3r_1+2\sqrt{2r_1^2+4}}{r_1+4}\right) \tag{16}$$

$$t(r_2) = -\frac{\nu}{4}\ln\left(\frac{3-4r_2}{3+4r_2}\right) \tag{17}$$

However, in most cases of sequence evolution the assumption that each position evolves with the same rate is inappropriate. This is especially true when considering coding sequences. For example as we have seen in figure 12 the sequences of the Polio gene VP1/2A showed almost completely identical first and second codon positions, and a largely randomized third codon position.

How can we detect whether or not in the diffusion model defined above the sequences evolve with uniform mutation rate or not? Assuming one uniform mutation rate a temporal reconstruction could be obtained from the mean Hamming distance. From a given mean Hamming distance the relative mutation time t can be obtained from equation (4). One parameter however can always be fitted. But we can immediately find out whether the assumption is correct. If the sequences mutate with one uniform rate then for given sequence space parameters $\overline{d_0}$, $\overline{d_1}$ and $\overline{d_2}$ and/or ratios r_1 and r_2 the two times $t(r_1)$ and $t(r_2)$ would have to coincide. If the two times differ (substantially) then the sequences mutate with at least two different rates.

We therefore want to generalise to a model in which there are r different typical substitution times t_i and in which $\nu_1,\ldots,\nu_r$ positions evolve with substitution times $t_1,\ldots,t_r$ respectively [12].

Then for a given set of parameters $\{\overline{d_0},\overline{d_1},\overline{d_2}\}$ the following system of equations has to be solved:

$$\begin{aligned} \nu_1+\nu_2+\cdots+\nu_r &= \nu \\ \overline{d_0}(t_1)+\overline{d_0}(t_2)+\cdots+\overline{d_0}(t_r) &= \overline{d_0}(t) \\ \overline{d_1}(t_1)+\overline{d_1}(t_2)+\cdots+\overline{d_1}(t_r) &= \overline{d_1}(t) \\ \overline{d_2}(t_1)+\overline{d_2}(t_2)+\cdots+\overline{d_2}(t_r) &= \overline{d_2}(t) \end{aligned} \tag{18}$$

And the following conditions have be fulfilled for each i:

$$t(r_1) = t(r_2)$$
$$\overline{d_0}(t_i)+\overline{d_1}(t_i)+\overline{d_2}(t_i) = \nu_i$$

This is a non-linear system of equations in the $3r$ unknowns $\overline{d_j(t_i)}$, $j=0,1,2$ and $i=1,\ldots,r$ for given values ν, $\overline{d_0}$, $\overline{d_1}$ and $\overline{d_2}$. In general it is only possible to solve these equations numerically. The following iterative procedure has been suggested [12]:

In the first step the value r is determined. Let $r=1$. Then the equation system (18) is automatically satisfied, so that one only needs to check whether $t(r_1)=t(r_2)$. If the equation is not satisfied, there must be more than one substitution rate.

Consequently r is set to 2 and the respective equations are solved. If that, too, proves impossible, r is set to 3 and so on.

5.1. *A model for the dating of diffusion processes exemplified for the AIDS-Virus*

The computation of the statistical geometry of the different HIV/SIV groups revealed a bundle-like topology for the common ancestor of the five groups (see fig. 15). This allows us, with the help of the analytical formulae developed in the last section, to estimate the age of the earliest node in this graph (see also [4] and [36]). In figure 17 a schematic graph of the HIV/SIV phylogeny is drawn. We assume that the sequences have evolved from a common ancestor

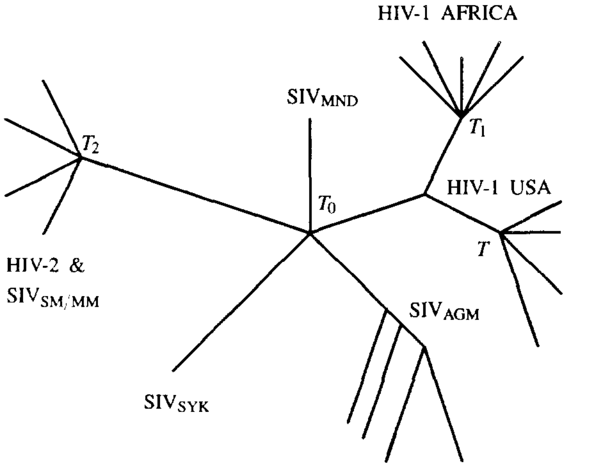

Fig. 17. Schematic graph of the phylogenetic relationships of the different HIV-SIV groups. From a given calibration time point T the divergence times T_0, T_1 and T_2 are to be determined.

at time T_0 independently and in parallel. At different times T_i new sequences have been generated through punctuation events. A calibration time is given by time T.

We first make the assumption that the sequences evolve with one uniform substitution time t. Then a temporal fit can be obtained from the mean Hamming distances within each group. For a given mean Hamming distance $\overline{d_{ij}}$ the relative mutation time t can be computed from equation (4). In table 2 the mean Hamming distances for two different genes are listed, as well as the resulting relative mutation time t of the four groups which form bundle-like graphs. The deduced absolute ages have been obtained assuming a calibration time T for the HIV-1 USA group of 15 years (reference time is 1985, the year of isolation for the sequences used).

From these computations it follows that the age of the common ancestor of all known five HIV/SIV groups is approximately 600 to 1100 years, a result also found by Li and co-workers [37].

In order to test the assumption of a uniform mutation substitution probability, the statistical sequence space parameters $\overline{d_0}$, $\overline{d_1}$ and $\overline{d_2}$ of the different groups were computed for both genes translated into the purine/pyrimidine sequences. From the two ratios $\overline{d_1}/\overline{d_0}$ and $\overline{d_2}/\overline{d_1}$ the relative mutation times were deduced according to equations (16) and (17). The values are plotted in figure 18.

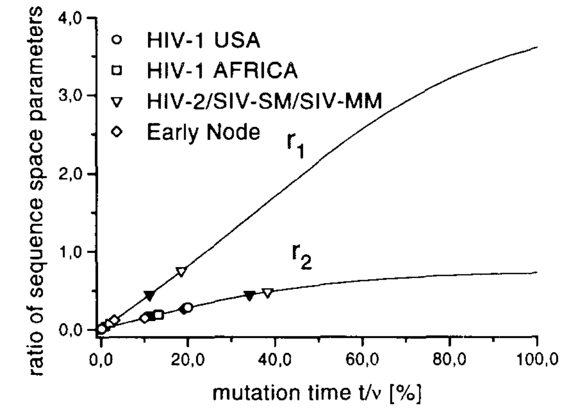

Fig. 18. Relative mutation times of the envelope gene and the core gene of different HIV/SIV groups, which were computed from the ratio $r_1 = \overline{d_1}/\overline{d_0}$ and $r_2 = \overline{d_2}/\overline{d_1}$ of the mean statistical geometry parameters using the equations (16) and (17). The open symbols represent the values of the envelope gene, the closed symbols those of the gag gene.

From the plotted values it is evident that in none of the cases the mutation times coincide. Furthermore, since the ratio r_2 is always larger than the ratio r_1 one can deduce that there is a large number of constant positions. However, the addition of a second typical substitution time does not suffice to fit the data. A minimum of three substitution times was needed to solve the equation system (18) unambiguously for all groups [12], [4]. The positions evolving with these different probabilities of substitution are categorized as constant, variable and hypervariable, thereby reflecting different resistances to the fixation of mutational change.

In table 3 the results of solving the equation system with three position categories are presented for both genes considered. From the deduced relative divergence times the absolute ages were determined using a reference time of 15 years for the HIV-1 USA group.

6. Computer implementation

Statistical geometry is naturally a method to be used with a computer. Therefore the author has written an easy-to-use, completely menu-driven program

Table 2
Relative mean Hamming distances of the listed groups of the envelope (ENV) and the core (GAG) gene with the resulting relative divergence times t and absolute ages.

Gene	*Group*	$\overline{d_h}/\nu$	t/ν	Ages [Years]
ENV	HIV-1 USA	0,005	0,003	15
	HIV-1 Africa	0,049	0,026	130
	HIV-2+SIV–MM/SM	0,097	0,054	270
	Earliest node	0,292	0,219	1095
GAG	HIV-1 USA	0,005	0,003	15
	HIV-1 Africa	0,025	0,013	65
	HIV-2+SIV–MM/SM	0,063	0,034	170
	Earliest node	0,217	0,136	680

Table 3
Temporal reconstruction of the common ancestor of different HIV/SIV groups, resulting from the relative divergence times (t_h, t_v, t_c) of the envelope (ENV) and core (GAG) sequences in the purine/pyrimidine space. Three categories of positions have been found (hypervariable ν_h, variable ν_v and constant ν_c). If the time of divergence of the HIV-1 USA group is 15 years, then the corresponding lower bound for the ancestor of the HIV-1 Africa group is obtained from t_h(Africa)/t_h(USA), for the ancestor of the HIV-2/SIV group from t_h(HIV-2)/t_h(Africa)+t_v(HIV-2)/t_v(Africa), and the lower bound for the earliest node is found from t_v(Earliest node)/t_v(HIV-2).

ENV gene: $\nu = 1896$ $\nu_h/\nu = 0.02$ $\nu_v/\nu = 0.73$ $\nu_c/\nu = 0.25$

	HIV-1 USA	HIV-1 Africa	HIV-2 & MM/SM	Earliest node
t_h/ν_h	0.20	0.39	0.69	∞
t_v/ν_v	0.00	0.04	0.08	0.374
Ages [years]	15	30	180	$\gg 800$

GAG gene: $\nu = 1230$ $\nu_h/\nu = 0.02$ $\nu_v/\nu = 0.56$ $\nu_c/\nu = 0.42$

	HIV-1 USA	HIV-1 Africa	HIV-2 & MM/SM	Earliest node
t_h/ν_h	0.15	0.32	0.66	∞
t_v/ν_v	0.00	0.02	0.08	0.327
Ages [years]	15	30	110	$\gg 500$

package, called STATGEOM, which is easily implemented on almost any computer system. STATGEOM allows the computation of the statistical geometry of any set of binary sequences, nucleotide sequences or amino acid sequences both in distance and sequence space. Extensive output as well as postscript graphs provide an interpretation and illustration of the results. STATGEOM is available free of charge under the following address:

http://www.mpibpc.gwdg.de/~kniesel.

7. Quadruple mapping

Recently a simple method, called likelihood mapping, for the visualization of the phylogenetic content of a sequence alignment has been developed [38]. The purpose of likelihood mapping is to display the tree-likeness of each quartet of a sequence alignment in a single graph. The values for the tree-likeness are obtained from the three maximum likelihoods belonging to the three possible fully resolved unrooted trees of quartets using any model for the evolution of the sequences. The three likelihoods are transformed into probabilities and the corresponding vector is mapped onto a two-dimensional simplex. The three corners of the simplex are the attractors of the three trees (see left simplex graph in figure 19a). Furthermore this simplex can be subdivided into seven areas, each being the basin of attraction of either one of the three fully resolved trees, or the region of the star phylogeny or the three rectangle geometries, where it is not possible to decide between two topologies (see right simplex graph in figure 19a).

Statistical geometry is easily applied to this procedure, and is accordingly coined quadruple mapping (Nieselt-Struwe and von Haeseler, in preparation). In contrast to the likelihood mapping method, the likelihood values are not used, but rather the values obtained from computing the sequence space parameters of each quartet. Since the degree of tree-likeness is decided on the basis of the box dimensions of the graph, we let x be equal to the support of the grouping of sequence A and B against sequence C and D, y be equal to the support of the grouping of sequence A and C against sequence B and D, and z be equal to the support of the grouping of sequence A and D against sequence B and C. We then define the vector $s = (s_1, s_2, s_3)$ where $s_1 = x/(x+y+z)$, $s_2 = y/(x+y+z)$, $s_3 = z/(x+y+z)$. s is mapped onto the two-dimensional simplex

$$S = \{s_1 e_1 + s_2 e_2 + s_3 e_3 | s_1 + s_2 + s_3 = 1, \\ 0 \leq s_1, s_2, s_3 \leq 1\}$$

where the e_i are the unit vectors. The computation of the basins of attraction is done exactly as in the likelihood mapping procedure.

In figure 20 we have shown the simplex graph from clustered HIV/SIV gag sequences used above. Though one tree seems to be favored, as indicated by the lower left simplex, the representation of all seven basins of attraction reveals that in fact all quartets form bundle-like topologies and are thus situated in the bundle attractor's area. This result stands in close accordance to the statistical geometries shown in figure 14.

For a set of n sequences there are $\binom{n}{4}$ possible quartets, if the sequences are not clustered. In this case we cannot distinguish seven areas in the trian-

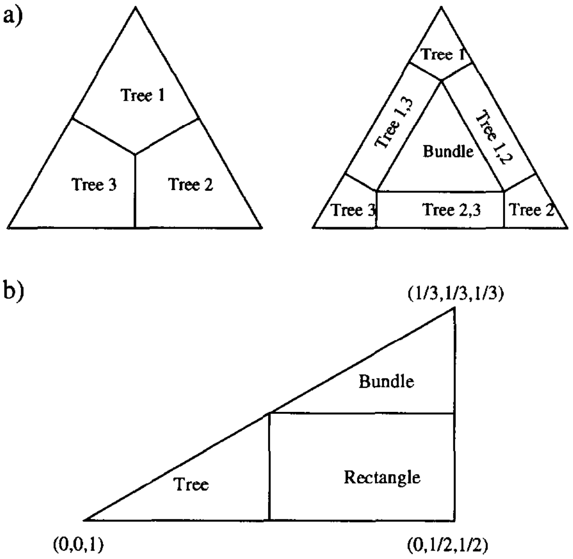

Fig. 19. (a) Mapping of the support vector $s = (s_1, s_2, s_3)$ onto the two-dimensional simplex using barycentric coordinates similar to the likelihood mapping procedure suggested by von Haeseler and Strimmer [38]. The corners in the upper left simplex represent the three fully resolved trees of four sequences with their respective basins of attraction. The upper right simplex shows the seven basins of attraction representing the three fully resolved trees, the bundle tree and the three regions of the rectangle topologies. (b) If unclustered sequences are used the complete simplex can be subdivided into six symmetric triangles. Each triangle has three basins of attraction representing the three global topologies: fully resolved tree, bundle or rectangle.

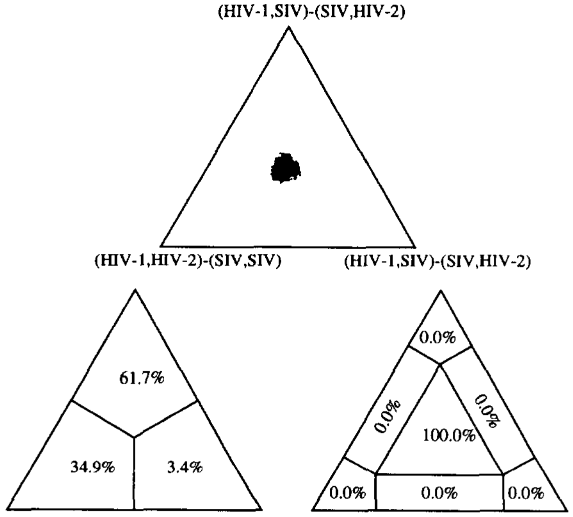

Fig. 20. Quadruple mapping simplex of the HIV/SIV sequences of figure 14.

gle, but only three: the area of tree-likeness, the area of the rectangle geometry and the area of bundle-likeness. This divides the simplex triangle into six symmetric triangles, each being fully equivalent to the others. We can therefore assume that the vector $s = (s_1, s_2, s_3)$ is ordered such that $s_1 \leq s_2 \leq s_3$. Then the corners of the representive triangle are given by the vectors $(0,0,1)$, $(1/3,1/3,1/3)$ and $(0,1/2,1/2)$. These are the attractors of the tree topology, the bundle topology and the rectangle topology respectively (see figure 19b). This mode of analysis yields an overall impression of the tree-likeness of the data.

8. Conclusions

When DNA or protein sequence data are analysed on the basis of some defined metric distance between different sequences, comparisons at the level of individual sequence positions are ignored. Distance-based analyses provide valuable information and can lead to a reasonably reliable reconstruction of the evolutionary history of the sequences, especially if independent point mutation has been the predominant process generating change. However, in the case that recombination events occur frequently, as is observed for viruses, the results of analyses based on distance methods often give a distorted view of the evolutionary history [39]. The method of statistical geometry allows data to be analysed concerning questions about the overall tree-likeness, tree-likeness in certain clusters, the degree of divergence globally or at individual positions. Statistical geometry is not meant to replace tree construction methods, but rather to provide an independent check on the validity of their implicit assumptions in particular circumstances. The results on viruses demonstrate the value of conducting such an evaluation. They show that important clues are obtained about the evolutionary history of the sequences and they also suggest changes to our previous view about the nature of viral diseases.

Similar conclusions can be drawn from an analysis of the application of statistical geometry to problems of non-biological configuration spaces such as the space of solutions to combinatorial optimization problems. Detailed investigations [12] (also, Nieselt-Struwe in preparation) reveals that only by using the method of statistical geometry in sequence space rather than distance-based methods can one discover that the local optima do not form a tree-like topology of the type usually assumed [40].

As more and more genetic sequence data become available, methods such as statistical geometry that allow to uncover sequence relationships reliably will become indispensable tools for every scientist interested in the reconstruction of the evolution of living beings.

Acknowledgements

It is a privilege to contribute to this special issue in honor of Manfred Eigen's 70th birthday. I wish to thank Manfred Eigen for introducing me to this field of the sciences, and for being my mentor of so many years. I am greatly indebted to Andreas Dress, who has lead me to the mathematical foundations of statistical geometry. I am grateful to the Deutsche Forschungsgemeinschaft and to the Alexander von Humboldt-Stiftung for support. I wish to thank Peter Wills and Peter Schuster for their critical suggestions on the manuscript.

References

[1] M. Eigen, R. Winkler-Oswatitsch, and A.W.M. Dress. Proc. Natl. Acad. Sci. USA, 85 (1988) 5913.

[2] M. Eigen and R. Winkler-Oswatitsch. Naturwissenschaften, 68 (1981) 217.

[3] M. Eigen and R. Winkler-Oswatitsch, Naturwissenschaften, 68 (1981) 282.

[4] M. Eigen and K. Nieselt-Struwe, AIDS, 4 (suppl. 1) (1991) S85.

[5] M. Eigen and P. Schuster, The Hypercycle — A Principle of Natural Self-Organization, (Springer Verlag, Berlin, 1979).

[6] S. Wright, in: Int. Proc. of the Sixth Int. Congress on Genetics, vol. 1, ed. D.F. Jones (Ithaca, New York, 1932) p. 356.

[7] R.W. Hamming, Bell Syst. Techn. J., 24 (1950) 147.

[8] R.W. Hamming, Coding and Information Theory, (Prentice Hall, Englewood Cliffs, 2nd edition, 1986).

[9] J. Maynard Smith, Nature, 225 (1970) 563.

[10] I. Rechenberg, Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Information, (Problemata, Formann-Holzboog, Stuttgart-Bad Cannstatt, 1973).

[11] R.F. Doolittle, Methods in Enzymology, 183, (1990).

[12] K. Nieselt-Struwe, PhD thesis, (Univ. Bielefeld, 1992).

[13] T.H. Jukes and C.R. Cantor, in: Mammalian Protein Metabolism, ed. H.N. Munro, (Academic Press, New York, 1969) p. 21.

[14] M. Kimura, J. Mol. Evol., 16 (1980) 111.

[15] K. Zaretskii, Upsheki Mat. Nauk, 20 (1965) 90.

[16] J.M.S. Simões Pereira, J. Comb. Theory, 6 (1969) 303.

[17] A.N. Patrinos and S.L. Hakimi, Quart. Appl. Math., 30 (1972) 255.

[18] H.-J. Bandelt and A.W.M. Dress, Adv. Math., 92 (1992) 47.

[19] M. Eigen and R. Winkler-Oswatitsch, Methods in Enzymology, 183 (1990) 505.

[20] R. Rico-Hesse, M.A. Pallansch, B.K. Nottay, and O.M. Kew, Virology, 160 (1987) 311.

[21] M.J. Rodrigo and J. Dopazo, J. Mol. Evol., 40 (1995) 362.

[22] J. Felsenstein, Evolution, 39 (1985) 783.

[23] R. Winkler-Oswatitsch, A. Dress, and M. Eigen, Chemica Scripta, 26B (1986) 59.

[24] M. Eigen, B. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A.W.M. Dress, and A. von Haeseler, Science, 244 (1989) 673.

[25] C.B. Mayer, PhD thesis, (Univ. Bielefeld, 1995).

[26] A. Rambaut and N.C. Grassly, CABIOS, (1997) in press.

[27] D.A. Buonagurio, S. Nakada, J.D. Parvin, M. Krystal, P. Palese, and W.M. Fitch, Science, 232 (1986) 980.

[28] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, Atlas of Protein Sequence and Structure, volume 5 suppl. 3, (Natl. Biomedical Research Foundation, Silver Spring, 1978).

[29] G.H. Gonnet, M.A. Cohen, and S.A. Benner, Science, 256 (1992) 1443.

[30] F.R. Schubert, K. Nieselt-Struwe, and P. Gruss, Proc. Natl. Acad. Sci. USA, 90 (1993) 143.

[31] J.W. Archie, Syst. Zool., 38 (1989) 219.

[32] D.P. Faith, Nature, 345 (1990) 293.

[33] D.P. Faith and P.S. Cranston, Cladistics, 7 (1991) 1.

[34] C. Kuiken, K. Nieselt-Struwe, G.F. Weiller, and J. Goudsmit, in: Methods in Molecular Genetics, Vol. 4, Molecular Virology Techniques Part A ed. K.W. Adolph, (Academic Press, San Diego, 1994) p. 100.

[35] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees, (Wadsworth & Brooks, Pacific Grove, California, 1984).

[36] K. Nieselt-Struwe and M. Eigen, in: Mathematik in der Praxis, (eds.) A. Bachem, M. Jünger and R. Schrader, (Springer-Verlag, Berlin, 1995) p. 291.

[37] W.-H. Li, M. Tanimura, and P.M. Sharp, Mol. Biol. Evol., 5 (1988) 313.

[38] K. Strimmer and A. von Haeseler, Proc. Natl. Acad. Sci. USA, (1997) to appear.

[39] J. Maynard Smith, TREE, 4 (1989) 302.

[40] S. Kirkpatrick and G. Toulouse, J. Physique, 46 (1985) 1277.