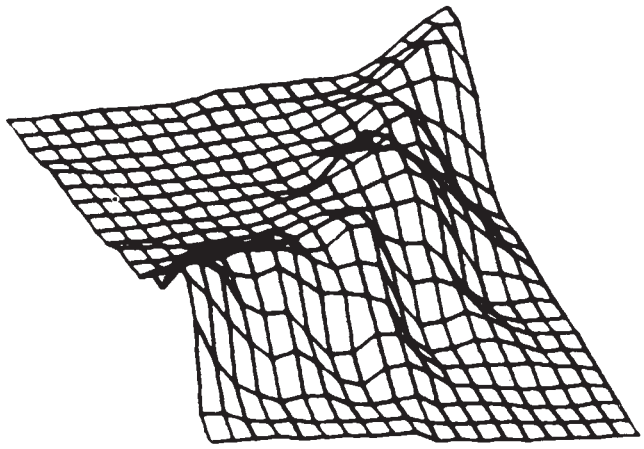

Fig. 1 A perspective grid derived from a matrix of values computed from linearized measurements.

across the lines. Measurements of many geophysical properties, made by moving vehicles, ships and so on, are often distributed along lines which are not necessarily straight. They may cross one another but the distance between adjacent measurements on the lines is usually much less than the distance between the lines. Such a distribution of measurement points may be found in other situations where the two dimensions are not spatial.

By Cole's method, the generalized surface is actually a matrix of Z values and the initial values of the matrix elements are derived from a global least squares fit of a second-order two-dimensional polynomial function. This global surface is then modified by a local least squares quadratic fit involving the 3 by 3 matrix around each data point. The modified matrix is then smoothed by means of an algorithm which removes discontinuities of gradient between the global and local surface but does not affect quadratic surfaces. The algorithm is made iterative by using the matrix obtained as the new global surface for a repeat of the process. The effect of this is the production of a smooth surface which passes through the data points. In areas where data are absent, the algorithm tends to extrapolate first and second derivatives. In practice this may be undesirable. Imagine a string of data points passing over a hill in the measured property. If the path deviates slightly from a straight line as it crosses the hill, this would be interpreted as a steep slope perpendicular to the track. But this would not be the subjective evaluation of the data, because such steep slopes are not found along the path of measurement. When the source data are linearized, more information is available about the statistical nature of the surface. I have therefore modified Cole's algorithm to improve these two aspects—the extrapolation of first and second derivatives, and the use of the statistical information. It is necessary to specialize somewhat because I propose that on a large scale neither the first nor second derivatives of the measured property can be extrapolated indefinitely. Instead, in any areas without data, we may expect a horizontal plane: this is true for many geophysical measurements but may not be valid for other types of property. Bearing this in mind, each element of the initial matrix of Cole is set equal to the value of the nearest data point. Then a single data point yields a single-valued plane, and at points away from the data the resultant surface tends to the value of the nearest data point. In this way improbable cross track gradients are not generated because initially the matrix values at either side of the track are equal to the value on the track, although there will in general be discontinuities between tracks. If the property being contoured has known discontinuities within the area of the matrix, in particular if there is a boundary (such as a known coastline) beyond which contours should not be plotted, then any data from beyond the boundary are omitted (the survey is discontinued at the boundary) and, after contouring, the plane surface which will extend beyond the boundary into the area of no data is omitted. The local quadratic least squares fit of Cole has been retained because on a small scale these extrapolations are permitted. The principal change to Cole's algorithm is in the smoothing. The source data are scanned along the survey lines and the root mean square gradient between adjacent data points is determined. We may now assume that the function which has been measured has statistical properties which are isotropic. To be more specific, I assume that the mean square gradient observed along the track is the same as that expected across the track. We may therefore smooth the matrix obtained by an amount which increases rapidly for gradients greater than the r.m.s. gradient determined along the track. The gradient in the vicinity of each matrix point is found, and the Z value changed by an amount $\Delta Z\, a/(1+a)$ in such a way as to reduce the gradient; a is the ratio between the local gradient and the r.m.s. measured gradient. Thus a matrix is produced representing a surface which passes through the data points and has the same statistical properties as the source data. This method is only possible because the data are distributed along lines, and it is not suitable for randomly scattered data. The matrix obtained is then contoured by a standard computer routine to produce a contour map. Alternatively, it is worth considering a perspective grid as shown in Fig. 1.

B. M. Ewen-Smith

British Antarctic Survey,
Scott Polar Research Institute,
Cambridge

Received May 17, 1971.

[1] Cole, A. J., *Nature*, **220**, 91 (1968).

Distance between Sets

Distance functions expressing the degree of dissimilarity of sets have found use in physical anthropology[1], psychology[2], numerical taxonomy[3], ecology[3] and elsewhere. During an ecological study by one of us, it was noticed that the similarity coefficient of Jaccard[6], used in ecology, gives rise to a metric function satisfying the triangle inequality. For two non-empty finite sets X, Y, the Jaccard coefficient is the number of elements in the intersection $X \cap Y$ of X and Y. This coefficient

$$r = \frac{|X \cap Y|}{|X \cup Y|}$$

(we use absolute value signs to indicate number of elements) has a heuristic interpretation. It measures the probability that an element of at least one of two sets is an element of both, and thus is a reasonable measure of similarity or "overlap" between the two. The one-complement

$$d(X, Y) = 1 - r(X, Y) \qquad (1)$$

may then be considered a measure of the dissimilarity of the two sets.

To study patterns of similarity or relatedness of sets of data, it is often helpful to represent them abstractly as points in a space, which can be studied by the methods of cluster analysis or ordination. For this, it is an advantage if the measure of dissimilarity has the formal properties of a metric or distance function.

If X, Y and Z are non-empty finite sets, we claim that

$$d(X, Y) + d(Y, Z) \geq d(X, Z) \qquad (2)$$

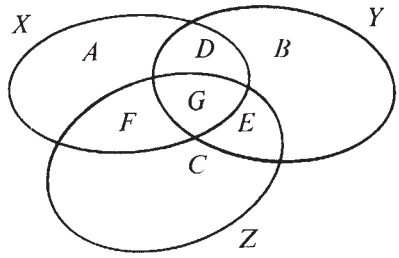

Fig. 1 Representation of sets and subsets.

By the definition of the function d, this reduces to showing that

$$\frac{|X\cap Y|}{|X\cup Y|}+\frac{|Y\cap Z|}{|Y\cup Z|}\leq\frac{|X\cap Z|}{|X\cup Z|}+1 \qquad (3)$$

We break up $X\cup Y\cup Z$ into seven disjoint subjects A to G which we use to transform inequality (3) into a manageable form. Thus

$$X\cup Y\cup Z=(A\cup B\cup C)\cup(D\cup E\cup F)\cup G$$

where

$$A=X-(Y\cup Z),\ B=Y-(X\cup Z),\ C=Z-(X\cup Y),$$
$$D=X-(A\cup Z),\ E=Y-(X\cup B),\ F=Z-(C\cup Y),$$
$$G=X-(D\cup F\cup A)=Y-(D\cup E\cup B)+Z-(E\cup F\cup C)$$

This is best understood from Fig. 1.

If $a, b, c, \ldots$ denote the number of elements in the corresponding sets $A, B, C, \ldots$, the inequality (3) becomes

$$\frac{d+g}{v-c}+\frac{e+g}{v-a}\leq\frac{f+g}{v-b}+1 \qquad (4)$$

where $v=a+b+c+d+e+f+g$, the number of elements of $X\cup Y\cup Z$. Symbolically written, the inequality (4) becomes

$$\frac{d'}{c'}+\frac{e'}{a'}\leq\frac{f'}{b'}+1$$

or

$$a'b'd'+b'c'e'\leq a'c'f'+a'b'c'$$

or

$$(v^2-av-bv+ab)d'+(v^2-bv-cv+bc)$$
$$e'\leq(v^2-av-cv+ac)f'+(v^2-av-bv+ab)c'$$

This inequality is equivalent to $0\leq v^3$

$$+v^2(-a-b-c-d-e+f-g)$$
$$+v[ab+ac+bc+ad+bd+be+ce+2bg-af-cf] \qquad (5)$$
$$+\{-abc-abd-abg-bce-bcg+acg+acf\}$$

The second term is equal to $v^2(2f-v)$ so that the sum of the first two terms is $2fv^2$. The inequality now, symbolically, is

$$0\leq v(2fv+[\ ])+\{\ \}$$

Because $v=a+b+c+d+e+f+g$ the two negative terms of [ ] cancel with terms in $2fv$. And the negative terms of { } cancel with terms in v[ ], namely with terms of the part

$$a(bc+bd+2bg)+b(ce)+c(2bg)+d0+e0+f0+g0$$

of v[ ]. Because all negative terms on the right side of inequality (5) cancel, the inequality is established.

$d(X, Y)$ satisfies the triangle inequality (2) and is also positive definite ($d(X, Y)=0$ if and only if $X=Y$) and symmetric ($d(X, Y)=d(Y, X)$), so it has the properties of a metric function in a space, the elements of which are finite non-empty sets. The argument presented does not depend essentially on the finiteness of the sets, and the theorem may be generalized by introduction of a suitable measure function for infinite sets.

We have not yet found a simpler or more elegant proof of this theorem based on the theorems of set-algebra. The absence of, or difficulty of finding, such a proof may explain the novelty of the result.

The function d is related to another metric function

$$d'(X,Y)=|X\cup Y-(X\cap Y)|$$

where $X\cup Y-(X\cap Y)$ is the symmetric difference of X, Y in that

$$d=\frac{1}{|X\cup Y|}d'$$

for non-empty finite sets X, Y, d', in turn, is a constant times the one-complement of a well-known measure of similarity, the simple matching function

$$r'=\frac{|X\cap Y|+|\overline{X\cup Y}|}{|X\cup Y|}$$

The properties of d and d' are, however, quite different, as one might expect from the probabilistic quality of d in ranging from 0 to 1. The probabilistic and geometric features induced by d are the subjects of further study.

We thank Howard Levene and F. E. Warburton of Columbia University for helpful advice.

MICHAEL LEVANDOWSKY

Haskins Laboratories,
Pace College,
41 *Park Row, New York, NY* 10038

DAVID WINTER

Department of Mathematics,
University of Michigan,
Ann Arbor, Michigan 48104

Received March 29; revised October 4, 1971.

[1] Mahanolobis, P. C., *Proc. Nat. Inst. Sci. India*, **2**, 49 (1936).
[2] McGill, W. J., *Psychometrika*, **19**, 97 (1954).
[3] Sokal, R. R., and Sneath, P. H., *Principles of Numerical Taxonomy* (W. H. Freeman, 1963).
[4] Orloci, L., *J. Ecol.*, **54**, 193 (1966).
[5] Levandowsky, M., thesis, Columbia University, New York, 1970.
[6] Jaccard, P., *Bull. Soc. Vaud. Sci. Nat.*, **38**, 69 (1902).

BIOLOGICAL SCIENCES

Specific RNA Methylase associated with Avian Myeloblastosis Virus

AT least three enzymes of DNA metabolites are found in the RNA tumour virus group—an RNA dependent DNA polymerase[1,2], a DNA dependent DNA polymerase[3] and an endonuclease[4]. We now report that a specific RNA methylase is associated with the avian myeloblastosis virus (AMV) which transfers a methyl group from S-adenosylmethionine to certain guanine bases in RNA to give N^2-methylguanine.

There is linear incorporation of the methyl group of S-adenosylmethionine by virus from a high-speed pellet which depends on the addition of both surfactant and tRNA (Fig. 1). There is, however, an appreciable time-dependent background, which does not represent incorporation into endogenous nucleic acid bases, nor does it seem to decrease when rate-zonal purified virus is compared with the high-speed pellet from viraemic plasma. Because this background represents almost one-third of the measured incorporation, we have devised a column assay to isolate and identify the major radioactive nucleic acid components formed in the reaction.

The major radioactive peak in the column assay consists solely of N^2-methylguanine as identified in six thin-layer chromatography solvents (including isopropanol–HCl–water,