LA-4973

CIC-14 REPORT COLLECTION
REPRODUCTION
COPY

c.3

Metrics in Biology, An Introduction

LOS ALAMOS NATIONAL LABORATORY
3 9338 00319 8339

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

UNITED STATES
ATOMIC ENERGY COMMISSION
CONTRACT W-7405-ENG. 36

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Printed in the United States of America. Available from
National Technical Information Service
U. S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22151
Price: Printed Copy $3.00; Microfiche $0.95

LA-4973
UC-32 & 48
ISSUED: August 1972

los alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87544

Metrics in Biology, An Introduction

by

W. A. Beyer
M. L. Stein
S. M. Ulam
Temple Smith*

*Northern Michigan University

LOS ALAMOS NATL. LAB. LIBS.
3 9338 00319 8339

METRICS IN BIOLOGY, AN INTRODUCTION

by

W. A. Beyer, M. L. Stein, S. M. Ulam

and

Temple Smith*
Northern Michigan University

ABSTRACT

The use of metrics in biology is discussed. Attention is given to metrics in pattern recognition, taxonomy, and especially molecular taxonomy. Various ways of constructing metrics between sequences for use in molecular taxonomy are discussed.

I. INTRODUCTION

To compare quantitatively different organisms, complex molecules, or biological entities in general, a measure of dissimilarity is required. More generally, all objects which form the elements of study in the natural sciences can be compared as to the degree of their differences. The notion of distance, in mathematics, is not directly or easily applicable to such studies, although intuitively any useful measure of the degree of difference between objects would seem to convey a measure of distance between them. The notion of distance between two sets of points in a metric space or between functions defined on some space (e.g., on the real line) is usually considered by comparing the values at each point separately. The differences are then either added in absolute value or integrated in the case of a continuum, and one may take instead of sums of absolute values, the square root of sums of squares of the differences, etc. This, however, refers to numeric objects, (sets, functions, operators) as rather fixed or rigid entities and does not in general involve moving or transforming one or both in order to obtain as close proximity or "fit" as possible. Obviously, if one wants to compare two given different organisms, even purely geometrical organisms only, one tries to place them in positions where the comparison is made with real "corresponding" parts.[1] Mathematically, this means that one looks for the distance between two sets, modulo a class of transformations. This class can, but need not necessarily, form a group.

In this report we shall recall the elementary notions about metric spaces; we shall then redefine distances and "pseudo distances" between sets or, what is particularly important, between those classes of sets defining the entities one deals with in the natural sciences. The stimulus for this general discussion stems in part from the interesting work of Fitch and Margoliash[2] and others[3-6] on reconstruction of evolutionary trees from the data on the amino acid sequence of certain proteins. It also comes from certain other unrelated studies of the general formulations of the problems of "recognition of patterns" and "artificial intelligence studies" which were undertaken by S. Ulam, R. Schrandt, and J. Mycielski.

In biological taxonomy there are attempts to define such mathematical concepts as subspaces and neighborhoods within the space of all organisms, yet the direct application of the more general concepts

*Supported in part by a National Institutes of Health postdoctoral fellowship, grant HD 42801.

of metric spaces to this apparent mathematical area of natural science has met with only limited success.[7] The work of Sokal and Sneath[8] exemplifies one of the more successful attempts to extract evolutionary measures or distances from numeric or phenotypic taxonomy. These studies were often concerned more with statistical analysis of the variables being analyzed than with more fundamental mathematical considerations.

The advent of modern molecular biology, and with it the availability of comparative protein sequence data, has renewed interest in numerical taxonomy on the part of both biologists and mathematicians.[9-11] This is largely because the data are simple enough to permit tractable evolutionary distance calculations, although the simple nature of the 23-element protein space or four-element DNA genetic space may be misleading. The biological interpretations of this new molecular taxonomy have raised controversies about our understanding of evolution.[12,13]

Because of the current interest in molecular taxonomy and morphological distances in general we have outlined in this report the mathematical concepts of metric spaces and distances which may be applicable to these areas. This outline is then followed by a discussion of the protein sequence problem.

II. DISSIMILARITY COEFFICIENTS AND METRICS

Let P be a set of objects. Following Jardine and Sibson (Ref. 14, pp.77-78), one says that a function ρ from P x P to the real line is a dissimilarity coefficient if it satisfies the following requirements:

1. $\rho(p,q) \geq 0$ for all $p,q \epsilon P$,
2. $\rho(p,p) = 0$ for all $p \epsilon P$,
3. $\rho(p,q) = \rho(q,p)$ for all $p \epsilon P$.

Sometimes one requires

4. $\rho(p,q) = 0$ implies $\rho(p,r) = \rho(q,r)$ for all $r \epsilon P$ (evenness) ,

or

5. $\rho(p,q) = 0$ implies $p = q$ for all $p,q \epsilon P$ (definiteness).

A dissimilarity coefficient which also satisfies

6. $\rho(p,r) \leq \rho(p,q) + \rho(q,r)$ for $p,q,r \epsilon P$ (triangle property)

in addition to properties 1-5, is called a metric. Intuitively, it would seem that any measure of distance should satisfy the triangle property. The triangle property is essential for relating meaningful topological notions to properties defined by distance. We should note here, however, that given any assignment of a "semidistance" satisfying only the first five properties one can obtain a metric from it by the following procedure: given two points, p,q, one considers all possible finite chains from p to q, continuing for example through $p, x_1, x_2, \ldots x_n, q$, and defining the distance from p to q as the minimum sum of the lengths of the chains:

$$\rho(p,q) = \min_{[n,x_1, \ldots, x_n]} \rho(p, x_1) + \rho(x_n,q) + \sum_{i=1}^{n-1} \rho(x_i, x_{i+1}) \quad . \qquad (1)$$

Sometimes it is useful to require that an additional property be satisfied:

7. $\rho(p,q) \leq \max(\rho(p,r), \rho(q,r))$ for all $p,q,r \epsilon P$ (ultrametric inequality).

The ultrametric inequality is important in the theory of p-adic numbers and valuation theory.[15] Its relevance to biology is brought out by the following.[14]

"The strongest assumption about evolutionary rates which can be made is that they are constant. On this assumption the dissimilarities between present-day populations would be monotone with the times since their divergence. They would therefore be ultrametric, since the times of divergence of populations in an evolutionary tree form an ultrametric. The fact that the dissimilarities between present-day populations are rarely ultrametric refutes the hypothesis of constancy of evolutionary rates in terms of known measures of dissimilarity."

The following geometric interpretation can be given to the ultrametric inequality: every triangle is isosceles and its base has length less than or equal to that of the equal sides.

III. METRICS IN THE SPACE OF CLOSED SETS, HAUSDORFF DISTANCE, AND APPLICATIONS

One of the more general metrics is the Hausdorff distance. (See Ref. 16, pp. 166-172; Ref. 17, pp. 214-224; and Ref. 18, pp. 20-32.) This is a definition of distance between closed sets of points in a metric space. We assume here compactness of the underlying space P (this means that given any sequence of points x_n of P there exists a subsequence of these points converting to some point of P). Given two closed sets, A and B, one defines:

$$\rho(A,B) = \underset{x\in A}{\text{Max}}\ \underset{y\in B}{\text{Min}}\ \rho(x,y) + \underset{y\in B}{\text{Max}}\ \underset{x\in A}{\text{Min}}\ \rho(x,y)\ . \tag{2}$$

This distance satisfies the triangle property. Under this definition of distance the class of all closed sets in P becomes itself a compact metric space that is denoted by 2^P. One can now iterate our definition and consider sets in the space 2^{2^P}. This means that we consider classes of sets. Again, one can define a distance between any two of these classes with the use of the Hausdorff formula. This will be important in the sequel because when we speak of properties of sets we really consider classes of sets having a given property. So, for example, when we speak of sets of points on a screen "looking" like a letter A, we mean the aggregate or the class of such sets, distinguishable from the class of sets which "looks" like a letter B. In this way when we define a distance between objects independently of their size and orientation, for example, we have to consider, given a set, the class of all sets obtained from it by translations and rotations and also by changing of scale; then, given two different objects we are led to two classes of sets. The degree of their similarity or a quantitative measure of their differences should take into account possible changes of scale and position. If we now consider these two classes of sets and take the Hausdorff distance between them we do in essence the following. Given a set of the first class we look at the set in the second class that is as close to it as possible; we then take a maximum of this with respect to all choices of the first set. We then perform it symmetrically the other way around by taking a set from the second class, etc. The sum (or the maximum) of these two numbers gives a measure of distance between the "letter A" and the "letter B."

In unpublished notes W. A. Beyer and S. Ulam have compiled possible methods of measuring distances between sets and certain theorems which should be proved about the distances.

IV. METRICS FOR MOLECULAR TAXONOMY

In molecular taxonomy one considers sequences of amino acids defining the same protein* in various species. This means, mathematically, a class of codes each consisting of a sequence of symbols or words with 23 possible symbols. The encoded information gives the physical, chemical, and structural properties of the protein. The length of the sentences for the protein cytochrome c, e.g., is about one hundred words, and the first task is to define a distance function between any two such sequences of symbols, each assuming values from 0 to 22.[2,4] This distance would then give a measure of dissimilarity. We shall, in this section, discuss the problem somewhat more generally. We may assume, for simplicity, that the symbols assume only two values: 0 and 1. We thus have a space of all sequences of this sort, of variable length, and we try to define a notion of distance between them under various postulates as to the equivalence or indistinguishability between some sequences. In other words, we shall assume, given a sequence, a class of other sequences "equivalent" to it and give definitions of distance between such classes. This is analogous with the illustration given above on classes of objects in the plane.

In mathematical studies, given two sequences of 0's and 1's, one may define the distance variously. For example, given $\alpha = [x_1, x_2, \ldots x_n]$ and $\beta = [y_1, y_2, \ldots y_n]$ as:

$$\rho(\alpha,\beta) = \sum_{i=1}^{n} |x_i - y_i| \quad , \tag{3}$$

or

*The phrase "same protein" means a class of proteins all of which perform the same biochemical function, and which are by implication evolutionarily related.

$$\rho(\alpha,\beta) = \sum_{i=1}^{n} (x_i - y_i)^2 \quad . \tag{4}$$

Still another way, suggested for coding problems by Hamming,[19] is:

$$\rho(\alpha,\beta) = \sum_{i=1}^{n} [1-\delta(x_i,y_i)] \quad , \tag{5}$$

where $\delta(x,y)$ is the Kronecker delta function. We note that this metric is equal to that defined by Eq. (3) only for binary sequences. This will be of value later when considering the problem of protein sequences which are formed from the 23-symbol amino acid space.

We might, however, assume that the given sequences of 0's and 1's are written not linearly but on a circumference of a circle, and we can arbitrarily rotate this circle rigidly so that each sequence of a given length n is equivalent to n-1 other sequences. Definition of distance, then, would concern a distance between classes of equivalents.

This is quite a general situation. In mathematics one defines a distance between two functions f(x) and g(x), for example, as follows:

$$\rho(f,g) = \text{Max}\ |f(x) - g(x)| \quad , \tag{6a}$$

$$\rho(f,g) = \int |f(x) - g(x)|\, dx \quad , \tag{6b}$$

or

$$\rho(f,g) = \left\{\int (f(x) - g(x))^2\, dx\right\}^{1/2} \quad , \tag{6c}$$

etc. We may, however, wish not to distinguish between functions which are obtained by shifting one from another. In this case we have to define distance between classes of functions, perhaps using the Hausdorff metric. Also consult the work of Marczewski and Steinhaus.[20]

We shall now consider still different definitions of distance between two sequences of 0's and 1's. One definition could be

$$\rho(\alpha,\beta) = \min_{n,\ m,\ n',\ m'} (n + m + n' + m') \tag{7a}$$

where n, m, n' and m' are defined by

$$(T_1)^n\ (T_2)^m\ \alpha = (T_1)^{n'}\ (T_2)^{m'}\ \beta \quad . \tag{7b}$$

Here we allow two types of transformation or two kinds of "steps." T_1 consists of changing a 0 into 1 or vice versa. T_2 consists of a deletion of a symbol anywhere in the sequence and subsequent contraction of the rest, to close the gap. Given two sequences, one may define as a distance the minimum total number of steps performed on one or both of these sequences so as to bring them into identical form. As an example, let α = [010101010101] and β = [101010101010]. Then by Eq. (3) or (5) we would obtain

$$\rho(\alpha,\beta) = 12 \tag{8a}$$

since all places have different values, while by Eq. (7) we would have

$$\rho(\alpha,\beta) = 2 \tag{8b}$$

since by deleting the first symbol in α and the last in β one obtains identical sequences.

If one considers mutations in a chain of DNA and if amino acids defining a protein are considered

to be of the special types defined above, then the distance as constructed above may correspond to the number of necessary mutations to transform one sequence into the other or both into an ancestral one. This mutational-transformation approach has been applied by Reichert and Wong[21] to protein and RNA sequences. Mathematically, this function of a pair of sequences is of certain combinatorial interest. It is not obvious a priori whether given at random two such sequences of n binaries the distance between them as computed by the algorithm above will be, on the average, a linear function of n. It is clear that this average will be less than n/2. It is also of interest to consider infinite sequences and diverse definitions of distances between them.

One can allow a number of transformations on sequences which lead from a given one to sequences which we consider equivalent among themselves. Given such a division into classes, one can define the distance function between classes à la Hausdorff indicated above, starting with a given notion of distance between individual sequences.

Another distance-type function which is applicable to the sequence problem can be defined as follows: given two sequences, we consider the number of 1's in each. We take the absolute value of the difference between them. Next we consider the number of 1's followed by 0's, compare that number in the two sequences, and take the difference. We do the same for 0's followed by 1's, then 1's followed by two 0's, etc. We then add the numbers. This type of "Markov distance" gives us perhaps an idea of the "visual" distance between the two given sequences.

The last two distances have considerable appeal for the molecular sequence problem. The sequence transformation metric defined by Eq. (7) would seem to have a direct biochemical interpretation, as pointed out above. However, for the nonbinary sequences (such as proteins defined in the 23-symbol amino acid space) the direct interpretations of the different transformation operators as the analogs of the physical mutations become more difficult. This is partly because the physical events take place in 4-symbol RNA space and are not always simple* functions of these noncommuting operators.[21]

*For example, genetic duplications, inversions, and the frame shifts as viewed in the protein space.

The "Markov distance" is also of considerable interest inasmuch as it appears to be a measure of the overall visual similarity of the sequences. This may be what is needed since in itself the sequence is not the object of ultimate interest but rather, as in the case of proteins, it is the three-dimensional structure which is, or chemical properties which are, encoded in the sequence.

It was in the light of the above considerations that a new sequence metric was defined. It began with an idea of Fitch.[9] Fitch's original proposal for detecting sequence homology was defined as follows. Let $\alpha = (X_1, \ldots, X_N)$ and $\beta = (Y_1, \ldots, Y_N)$ be two sequences of amino acids each of length N. Let $\eta(X,Y)$ be some measure of the distance from amino acid X to amino acid Y. Put

$$\rho_n(\alpha,\beta) = \sum_{\ell=0}^{N-n} \sum_{k=0}^{N-n} \sum_{j=\ell+1}^{n+\ell+1} \eta(X_j, Y_{k+j}) - np(N-n+1)^2, \quad (1 \le n \le N) \quad . \tag{9}$$

Here n is what is thought to be a statistically important subsequence size. The second term is the expected value of the measure assuming nonhomologous or random sequences with an average element probability of p. Our new sequence metric can be defined in a related manner. Put

$$\gamma(\alpha,\beta,n) = \sum_{\ell=0}^{N-n} \min_{0\le k\le N-n} \sum_{j=\ell+1}^{n+\ell+1} \eta(X_j, Y_{k+j-\ell}) \tag{10a}$$

and

$$\rho'(\alpha,\beta) = \left[\sum_{n=1}^{N} \left(\frac{\gamma(\alpha,\beta,n)}{N-n+1}\right)^2\right]^{1/2} \left[\sum_{n=1}^{N} n^2\right]^{-1/2} . \tag{10b}$$

Then the metric is

$$\rho(\alpha,\beta) = \mathrm{Max}\left\{\rho'(\alpha,\beta);\ \rho'(\beta,\alpha)\right\} \tag{10c}$$

This metric has a number of potential advantages including the fact that it can be applied to sequences of varying length although no proof exists as to the triangle inequality for such cases. It also can give a measure as to the degree of redundance (subsequence duplications within or among the sequences). In a subsequent[22] paper we shall study this metric in greater detail.

V. REMARKS

a. One of the major uses of distances in biology is in cluster analysis and evolutionary tree construction. J. A. Hartigan (unpublished notes) has pointed out the following objection: pairwise distances are a more sophisticated form of dissimilarity judgment than clusters, and so it may be inappropriate to use them to compute clusters. However, in tree construction where one wants to estimate length of branches, the distance concept is useful. There are other situations where one wants an estimate of the distance between clusters, and the distance concept is useful.

b. Another quantity which might be used in place of distance is a quantity σ to be thought of as relating to the probability of transition from p to q. σ (p,q) should be a mapping from P x P to [0,1] which satisfies

1. $0 \leq \sigma(p,q) \leq 1$
2. $\sigma(p,q) \geq \sigma(p,r)\sigma(r,q)$

for all $p,q,r \in P$. Development of a theory of such a function might be worthwhile.

REFERENCES

1. E. S. Smirnov, "Mathematische Studien uber Individuelle und Kongregationen Variabilität," Verh. 5 Intern. Kong. Vererbungswiss. 2, 1373-1392 (1927).
2. W. M. Fitch and E. Margoliash, "Construction of Phylogenetic Trees," Science 155, 279-284 (1967).
3. T. H. Jukes, Molecules and Evolution (Columbia Univ. Press, New York, 1966).
4. M. O. Dayhoff, Atlas of Protein Sequence and Structure (National Biomedical Research Foundation, Silver Spring, Md., 1969).
5. D. E. Kohne, "Evolution of Higher Organisms DNA," Quant. Rev. Biophys. 3, 327-381 (1970).
6. M. Goodman, J. Barrabas. G. Matsuda, and G. W. Moore, "Molecular Evolution and the Descent of Man," Nature 233, 604-613 (1971).
7. E. Mayr, Principles of Systematic Zoology (McGraw-Hill, New York, 1969).
8. R. R. Sokal and P. H. A. Sneath, Principles of Numerical Taxonomy (W. H. Freeman and Co., San Francisco, 1963).
9. W. M. Fitch, "An Improved Method of Testing for Evolutionary Homology," J. Mol. Biol. 16, 9-16 (1966).
10. T. Uzzell and K. W. Corbin, "Fitting Discrete Probability Distributions to Evolutionary Events," Science 172, 1089-1096 (1971).
11. M. Kimura, "Evolutionary Rate at the Molecular Level," Nature 217, 624-626 (1968).
12. B. Clarke, "Selective Constraints on Amino-Acid Substitutions During the Evolution of Proteins," Nature 228, 159-160 (1970).
13. T. H. Jukes and J. L. King, "Deleterious Mutations and Neutral Substitutions," Nature 231, 114-115 (1971).
14. N. Jardine and R. Sibson, Mathematical Taxonomy (John Wiley and Sons, New York, 1971).
15. G. Bachman, Introduction to p-adic Numbers and Valuation Theory (Academic Press, New York, 1964).
16. F. Hausdorff, Set Theory (Chelsea Publishing Co., New York, 1957).
17. K. Kuratowski, Topology, Volume I (Academic Press, New York, 1966).
18. K. Kuratowski, Topologie, Volume II (Polish Scientific Publishers, Warsaw, 1961).
19. R. W. Hamming, "Error Detecting and Error Correcting Codes," Bell System Tech. J. 29, 147-160 (1950).
20. E. Marczewski and H. Steinhaus, "On a Certain Distance of Sets and the Corresponding Distance of Functions," Colloq. Mathematicum 6, 319-327 (1958).
21. T. A. Reichert and A. K. C. Wong, "An Application of Information Theory to Genetic Mutations and Matching of Polypeptide Sequences," to appear in J. Theor. Biol.
22. W. A. Beyer, T. Smith, M. L. Stein, and S. M. Ulam, "A Molecular Sequence Metric and Evolutionary Trees," submitted to Nature.

HK/rr:499/200