J. theor. Biol. (2001) **212,** 57–69
doi:10.1006/jtbi.2001.2347, available online at http://www.idealibrary.com on IDEAL®

The Geometry of Shape Space: Application to Influenza

ALAN LAPEDES*†‡ AND ROBERT FARBER‡*

* *Theoretical Division, MS B*213, *Los Alamos National Laboratory, Los Alamos, NM* 87545, *U.S.A. and*
‡ *The Santa Fe Institute*, 1399 *Hyde Park Road, Santa Fe, NM* 87501, *U.S.A.*

(*Received on* 13 *December* 2000; *Accepted in revised form on* 3 *May* 2001)

Shape space was proposed over 20 years ago as a conceptual formalism in which to represent antibody/antigen binding. It has since played a key role in computational immunology. Antigens and antibodies are considered to be points in an abstract "shape space", where coordinates of points in this space represent generalized physico-chemical properties associated with various (unspecified) physical properties related to binding, such as geometric shape, hydrophobicity, charge, etc. Distances in shape space between points representing antibodies and (the shape complement) of antigens are assumed to be related to their affinity, with small distances corresponding to high affinity.

In this paper, we provide algorithms, related to metric and ordinal multidimensional scaling algorithms first developed in the mathematical psychology literature, which construct explicit, quantitative coordinates for points in shape space given experimental data such as hemagglutination inhibition assays, or other general affinity assays. Previously, such coordinates had been conceptual constructs and totally implicit. The dimension of shape space deduced from hemagglutination inhibition assays for influenza is low, approximately five dimensional.

The deduction of the explicit geometry of shape space given experimental affinity data provides new ways to quantify the similarity of antibodies to antibodies, antigens to antigens, and the affinity of antigens to antibodies. This has potential utility in, e.g. strain selection decisions for annual influenza vaccines, among other applications. The analysis techniques presented here are not restricted to the analysis of antibody–antigen interactions and are generally applicable to affinity data resulting from binding assays.

© 2001 Academic Press

1. Introduction

"Shape space" was introduced by Edelstein and Rosen (Edelstein *et al.*, 1978), and Perelson and Oster (Perelson *et al.*, 1979) as a conceptual and computational framework in which to view antibody-antigen affinity and its resultant consequences. It has since played an important role in theoretical and computational studies of the immune system (Segel & Perelson, 1988; Perelson, 1988; DeBoer *et al.*, 1992, 1992a). This paper presents algorithms related to ordinal and metric multidimensional scaling (Shepherd, 1963, 1964; Kruskal, 1964a, b) which create an *explicit* representation of shape space and which provide numerical coordinates, given suitable experimental data, which represent molecular positions in the space. Although the numerical precision of affinity measurements is often limited, the algorithms described here are robust, and can construct quantitative information such as numerical coordinates, from qualitative information such as

† Author to whom correspondence should be addressed. E-mail: asl@lanl.gov

0022–5193/01/170057 + 13 $35.00/0
© 2001 Academic Press

the rank order of affinities as provided by a panel of experimental data, e.g. hemagglutination inhibition (HI) assay results. Hemagglutination inhibition assays measure the ability of antibodies to bind to antigens. In the context of influenza, the assay reports the ability of ferret antibodies raised against one viral strain to inhibit a second strain's ability to agglutinate red blood cells. If attempts are made to define similarity of antigens, respectively antibodies, using binding assay data without reconstructing the geometry of the underlying shape space, then significant errors can result as we demonstrate below.

The idea of shape space as originally developed in the context of antibody/antigen binding is simple yet powerful, and presumably applies to other molecular interactions. Here we concentrate on antibody/antigen interactions. Each antibody and each antigen is assumed to be implicitly described by a vector of numbers, i.e. a coordinate vector, which represent the geometric shape characteristics relevant to shape complementarity in binding, as well as more general physico-chemical characteristics related to binding. These shape and physico-chemical characteristics need not be known for any individual molecule, but are assumed to exist, and are assumed to be sufficient to provide a complete description of molecular binding if they were known. Each vector represents an antibody, respectively antigen, as a point in a generalized "shape space" of some (to be determined) dimension. Antigens which are bound tightly by an antibody are assumed to have similar shape space vectors (or more precisely, similar complement shape vectors, see below) to the antibody, and hence are described by points in shape space which are close in Euclidean distance as calculated from their coordinate vectors to the antibody point. Experimentally observed affinity values are assumed to be a monotonic transformation of the distance between an antibody and an antigen in the underlying shape space.

In previous work (Perelson *et al.*, 1979; Segel & Perelson, 1988; Perelson, 1988; DeBoer *et al.*, 1992, 1992a) these coordinate vectors remained as implicit theoretical constructs, but even though implicit, the shape space formalism provided a powerful conceptual framework in which to explore molecular affinity and related issues. In this work, we provide algorithms which calculate explicit coordinate vectors in shape space from experimental data, and provide a formalism in which to execute quantitative investigations. A point requiring mention is that, in general, complementary shapes bind well (or more precisely, complementary physico-chemical characteristics lead to good binding), and hence the shape space vector describing one of the members of the pair (antibody, bound antigen) actually describes the complementary "shape" for that member. The word "shape" in this context denotes geometric as well as other physico-chemical characteristics of molecular surfaces relevant to binding, and does not necessarily imply a "lock and key" concept of molecular affinity.

Perelson and Oster were able to estimate certain gross properties of shape space, such as bounds on the dimension of the space, from experimental data even though they had no means to assign actual coordinate vectors in shape space to molecules. The dimension estimated by Perelson and Oster turned out to be fairly low (approximately five-dimensional), a value validated via our quantitative analysis using independent methods on different experimental data. A key contribution of this paper is the development and application of algorithms which provide an estimate of dimension, as well as explicit coordinates for molecules in shape space, given experimental data such as hemagglutination inhibition (HI) assays. Both the dimension, and coordinates, are determined by minimizing an objective function which relates the coordinates of points in shape space, to the given binding data, via a monotonic map from shape space distance to measured affinities. The algorithms are robust, and even though assay data is typically of low precision, the algorithms can produce high-quality coordinates which provide a detailed description of the geometry of shape space given only low precision experimental data. This recovery of high precision metric information from low precision data is a characteristic of the class of algorithms known as ordinal multidimensional scaling algorithms (Borg, 1997), to which our work is closely related.

Our formalism provides a quantitative description not only of the binding of antigen to antibody, but also allows one to compute measures

of similarity of one antigen to another antigen, and of one antibody to another antibody. Various applications of the formalism exist, including analysis of hemagglutination inhibition (HI) assay data used, e.g. in selection decisions for components of the annual influenza vaccine. We present results of analyses of various HI assay panels.

2. Material and Methods

2.1. MEASURING SIMILARITY: DIFFICULTIES IN CONVENTIONAL APPROACH

It is possible to calculate distances, i.e. (dis)similarities, between either the antibodies, considered as a set, or the antigens considered as a distinct set (but not between an antibody and an antigen) by defining coordinates for either the antibodies or the antigens in the following simple manner: view the experimental values for M antigens binding N antibodies as an $M*N$ panel of numbers, and consider the rows to be coordinates for the antigens or, respectively, consider the columns to be coordinates for the antibodies in a Euclidean space. The rows, respectively columns, can be thought of as "feature vectors" describing either the antigens or the antibodies. A distance, or (dis)similarity between antigens, respectively antibodies, can then be defined using the standard (or possibly weighted) Euclidean distance between feature vectors. By construction, antibodies and antigens are not represented in the same space, and furthermore the dimension of the resulting space is the (arbitrary) number of antibodies, respectively, antigens, in the panel.

This intuitive method of defining similarities can be misleading when affinities are in fact related to distances in an underlying space of fixed dimension, as per shape space assumptions. A simple simulation of shape space demonstrates this. Ten points representing antibodies are scattered into a space of five dimensions by choosing coordinates for the ten points at random between -1 and $+1$ in each of the five dimensions. Hence, N of a simulated NM panel is $N = 10$. Next, an additional 50 points are scattered in the space in a similar fashion to represent 50 antigens in the five-dimensional shape space. Hence, $M = 50$. The fact that experimental panels do not usually have e.g. 50 antigens is irrelevant. We use 50 antigens in this $50 * 10$ example merely to make trends visually apparent in Fig. 1. Compare distances between antigens (one could equally consider antisera) as calculated in this ten-dimensional space, to the distances between antigens as calculated in the true underlying five-dimensional shape space.

As can be seen in Fig. 1, attempting to select similar points based on the ten-dimensional "panel distance", by horizontally slicing in the y value, results in a wide range of associated distances in the true five-dimensional space represented by the x value. Hence, for all but the very smallest y values, the simplistic procedure to define similarities can result in a wide range of similarity values in the true space, including quite poor similarities.

2.2. SIMILARITY MEASURES FROM AFFINITY PANELS: A SHAPE SPACE APPROACH

"Multidimensional scaling" algorithms (referred to hereafter as MDS) are a class of algorithms initially developed in the computational psychology literature (Shepherd, 1963, 1964) which reconstruct the true dimension of the space, and the relative coordinates of points, given only distances, or more generally monotonic transformations of distances, between the points. "Relative" means that coordinates are reconstructed from the distance data up to global translation, reflection, scale and rotation which leave the relative relation of points invariant. Since interest centers on relative relationships, such global transformations are irrelevant.

There are two classes of MDS algorithms: (1) "metric MDS' algorithms, for which distances between points are given as input to the algorithms, and (2) "ordinal MDS" algorithms, for which the rank order of distances are given as input (Kruskal, 1964a, 1964b). Metric MDS has seen previous use in biological applications (Braun, 1987). Ordinal MDS algorithms have been used extensively in the computational psychology literature to derive quantitative conclusions from qualitative data, such as a human subject's relative rankings of the visual similarities of pairs of objects (Edelman, 1995). Somewhat surprisingly, if only the rank order of a set of

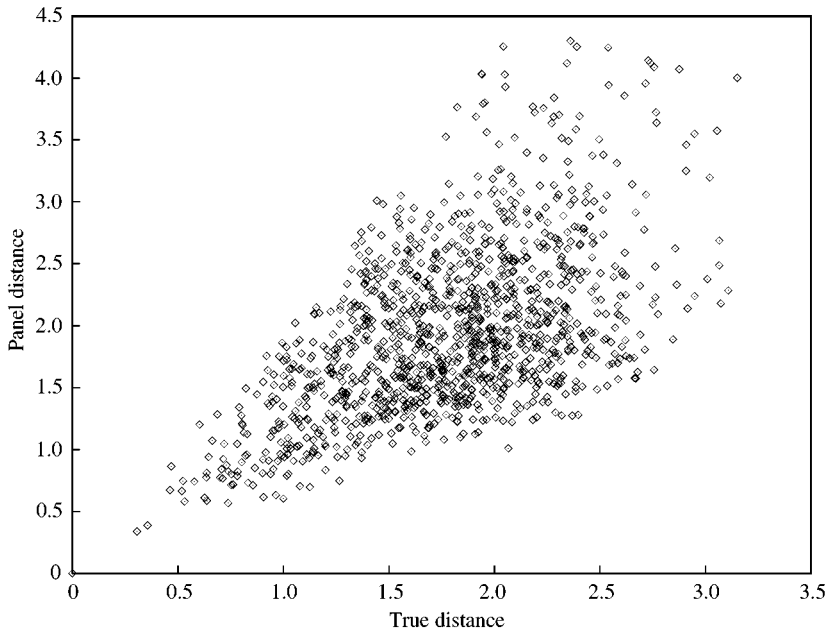

FIG. 1. Scatter plot for antigens with true shape space distance on the x-axis, vs. panel distance on the y-axis, for a panel of 50 antigens and ten antibodies in a five-dimensional shape space. A small distance corresponds to high similarity. Horizontal slices in y-value (i.e. fixed "panel distance") are associated with a range of x values (i.e. true distances), and even relatively small panel distances can have large x values (true distances), demonstrating that panel-based similarity measures typically include points with questionable similarity in the true space. More precise methods, such as described in this paper, are required for an accurate similarity determination.

distances between points are known, it is still often possible to compute with a high degree of precision the coordinates of the points in shape space giving rise to the ranked set of distances (Edelman, 1995). This is because given enough data, the set of rank relations impose sufficient inequalities on distances between points in the space such that the resulting set of simultaneous inequalities slices the space into small allowable regions in which each point can exist. The lower the dimension of the space, the less data are required for recovery of metric information from ordinal data. Fortunately, the dimension of immunological shape space turns out to be low.

Following Perelson and Oster (Perelson *et al*., 1979) we assume that affinity measurements between antibodies and antigens are described by an *a priori* unknown monotonic transformation of the distances between the points describing them in shape space. The ability to work with rank order data encompasses totally general monotonic transformations between distances in shape space and experimentally measured values such as hemagglutination inhibition assays, or more general binding assays. For shape space problems considered here, one is given experimental measurements (assumed to be monotonically related to distances), involving a subset of the data: only antigen–antibody measurements are given, and the antibody–antibody and antigen–antigen distances are then calculable after the algorithm reconstructs coordinates in the true underlying shape. Ordinal MDS applied to a subset of the data is known as the "unfolding problem" in the MDS literature (Borg, 1997).

2.3. ALGORITHMS

2.3.1. *Metric MDS*

For pedagogical purposes we first assume that experimental measurements are in fact distances, and not more generally, monotonic transformations of distance. Ordinal MDS will later be used to address experimental values which are monotonic transformations of distance. The computational task of metric MDS may be formulated as minimizing the following objective function as a function of coordinates

$$E = \sum_{i,j=1}^{i,j=M,N} (D_{ij}^{expt} - D_{ij})^2,$$

where D_{ij}^{expt} are the known, experimentally determined, "distances" (or more generally, monotonic transformations of affinity measurements), and the estimated Euclidean distances D_{ij}, are computed in a standard fashion given the coordinate vectors, X, of the points in a space of dimensionality D (D to be determined). Hence, $D_{ij} = |X_i - X_j|^2$, where $||^2$ represents the usual vector norm. The D-dimensional vector, X_i, represents the position of the i-th of M antigens in shape space, and similarly X_j represents the position of the j-th of N antibodies (or more generally, antisera). Each vector, X_i, has components denoted as superscripts, $X_i = (X_i^1 X_i^2 X_i^3 \cdots X_i^D)$, where the numerical values of the components, and the dimension D, are to be determined. Various weights can be introduced into E, if desired, to weight smaller distances (higher HI values) more heavily. The definition of D_{ij}, above, assumes a Euclidean reconstructed shape space. If shape space were in fact non-Euclidean, but reconstruction of shape space was based, as in the above, on a Euclidean assumption, then an artificially inflated dimension will result. A simple example of this is the embedding of the intrinsically two-dimensional, curved, surface of a hemisphere in a minimum of three Euclidean dimensions. Given the results obtained to date and described below, consideration of a more complicated non-Euclidean shape space reconstruction algorithm does not seem necessary.

One may minimize E as a function of the components using, e.g. steepest descents or conjugate gradient methods, to find optimal coordinates. Clearly, a trivial minimum is always possible. This occurs when one chooses the dimension, D, of the space to be greater than or equal to $P - 1$ where P is the number of points. Because P points can always be embedded in a Euclidean space of $P - 1$ dimensions, e.g. three points define a plane in two dimensions, a non-trivial embedding is obtained only if the objective function can be minimized when D is substantially less than P. One can therefore test possible embedding dimensions, starting with a low dimension, and plotting the final value of the objective function, E, as a function of the trial embedding dimension. A non-trivial embedding will be evidenced by a significant reduction of the final value of the objective function at a dimension, $D = D_{minimal}$, which is considerably less than P, the number of embedded points. Satisfying the condition, $D_{minimal} \ll P$ defines a low-dimensional submanifold constituting a non-trivial Euclidean embedding. Ordinal MDS, described below, uses a different objective function to define a non-trivial Euclidean embedding.

The difficulty with applying metric MDS to shape space is that the monotonic transformation relating HI values to shape space distances, D_{ij}^{expt}, is *a priori* unknown. It may be verified via simulated examples (data not shown) that if a monotonic transformation is used to reconstruct shape space which does not match the actual transformation between HI and distance, then (a) the reconstructed dimension is typically artificially inflated, and (b) a spread of reconstructed distances against true distances (similar to Fig. 1) is obtained. Ordinal MDS, considered next, avoids these problems.

2.3.2. *Ordinal MDS*

Ordinal MDS addresses the monotonic transformation problem by using only rank order information. The numerical values of the experimental data are not used other than to sort the values. This encompasses arbitrary monotonic transformations, since such transformations leave the rank order invariant. If there are M antigens and N antibodies then there are $M*N$ experimental values, E_{ij}, to be related to the distances between the representative points in shape space. Each E_{ij} is assumed to be monotonically related to the distance D_{ij} in shape space between antigen i and antisera j via an *a priori* unspecified monotonic function.

The E_{ij} values can be ordered by simple sorting from high to low. Initially, coordinates for the $M*N$ points are chosen at random, and so the ordering defined by sorting the E_{ij} values will not necessarily agree with the ordering defined by sorting the associated D_{ij} values. The algorithm seeks to move coordinates of points so that ultimately these orderings agree in the sense that HI values increase as shape space distances decrease. To achieve this, index the sorted list of E_{ij} values and the associated D_{ij} with a number α from 1 to MN, such that rank one is the highest experimental value which ultimately is to be associated

with the smallest distance. For example, if the experimentally determined HI value for the pair ij is the highest among all the antigen–antibody pairs then pair ij is assigned an ordering label, α, with $\alpha = 1$. We then denote the distance D_{ij} between ij as D_1, i.e. D_α with $\alpha = 1$. If the next highest HI value is between pair kl then the pair kl is assigned $\alpha = 2$, and the distance D_{kl} between kl is denoted D_2, i.e. D_α with $\alpha = 2$. We wish to find coordinates of all the points such that $D_1 = D_{ij}$ is less than $D_2 = D_{kl}$, and similarly and simultaneously, $D_2 < D_3$, $D_3 < D_4$, etc. In other words, we wish to find coordinates such that the computed distances are in correct association with the experimental values, i.e. high HI values correspond to small shape space distances.

An objective function, which when minimized as a function of shape space coordinates ranks the computed distances in shape space in the desired order in relation to the experimental values, is

$$E = -\sum_{\alpha=1}^{\alpha=MN} \log(g(D_{\alpha+1} - D_\alpha)), \qquad (1)$$

D_α references the MN computed distances using the index, α, based on the rank ordering of the experimental values explained above. $g(x)$ is a sigmoidal function which is zero at large negative values of its argument and one at large positive values e.g. $g(x) = 0.5 * (1 + \tanh(x))$. The exact algebraic form is not critical, and it is stressed that this monotone function is not at all related to the monotonic function relating distances in shape space to HI values.

Note that when the rank order of the computed distances is in the desired relation to the HI values (which occurs when coordinates are found such that $D_1 < D_2 < D_3 \cdots < D_{MN}$) then the $g(\,)$ function of each term of eqn (1) tends towards the value 1 and hence $\log(g(\,))$ tends towards the value 0. Thus, this objective function is minimized as a function of the antigen/antisera coordinates in shape space, achieving a value equal to zero, when the rank order of the computed distances is in desired relation to the experimental values. A non-trivial embedding is obtained when the desired rank order relation can be obtained in dimension $D = D_{minimal}$ considerably less than NM, the number of points. It is a somewhat suprising, but classic result of ordinal MDS theory (Borg, 1997), that this rank order restriction in fact places great restrictions on the possible coordinates of points, if $D_{minimal}$ is considerably less than the number of points.

Conjugate gradient algorithms work well in implementing an efficient minimization. Local minima, a minor problem in the analyses described below, can easily be surmounted by choosing a few initial starting values for the coordinates of the points. Simultaneous coordinates are produced for antigens and antibodies, hence antigen–antigen and antibody–antibody distances are defined. The distances between antigens and antibodies are, by construction, related to their affinities. Antigen to antigen distances and antibody to antibody distances quantify the similarity among antigens, respectively antibodies, with regard to the interaction(s) being measured.

3. Results

Influenza is a rapidly mutating RNA virus for which there is a national and international policy of annual vaccination. Abundant HI data are produced each year to assess the cross-reactivity of different annual strains of influenza with antisera that has been raised against strains of interest, typically those of preceding years. Such data are an important component of a decision process involving HI data, sequence data, and epidemiological data to select strains for inclusion in the influenza vaccine for any given year. A more detailed analysis of the antigenicity and evolution of influenza virus using our methods will be given elsewhere. Here we concentrate on the determination of the dimension of immunological shape space given experimental HI data, and the exposition and validation of the algorithms which reconstruct shape space from experimental data.

We apply our version of ordinal multidimensional scaling to various data sets of HI values for influenza below. These comprise two published panels of HI values (Raymond *et al.*, 1986; Both *et al.*, 1983), five unpublished HI panels (Centers for Disease Control, CDC, priv. comm.) representing repeated determinations on five separate days of HI values for identical sets of antigens

and antisera (to test invariance of the recovered geometry to experimental uncertainties in the determination of HI values), and finally, various sets of simulated HI data (to test if special properties of HI tables, e.g. the fact that they result from two-fold dilution studies, can produce artificially low dimensions under MDS analysis). All data sets obtained from laboratory experiments turn out to have a dimension equal to either four or five.

3.1. ANALYSIS OF HI ASSAY DATA: H1 INFLUENZA HEMAGGLUTININ 1950–1957 AND 1977–1983

The first data set we consider consists of a panel of HI values for 19 antigens (influenza viral strains) vs. 14 antisera (hence, $M = 19$ and $N = 14$) for influenza H1 subtype hemagglutinin, published in an investigation by Raymond *et al.* (1986). The 19 antigens are selected strains of influenza virus from the years 1950–1957 and 1977–1983, the 14 antisera are corresponding antisera to a subset of these 19 antigens.

Application of eqn (1) results in the rank order of the hemagglutination inhibition (HI) assay values being preserved in a minimum of five dimensions. A clear signature, illustrated in Figs 2 and 3, of the underlying dimension of shape space is given by the minimal dimension in which points representing antisera and antigens may be embedded without rank errors. The number of rank errors is defined to be the number of times the inequality $D_{\alpha+1} > D_\alpha$ of eqn (1) is violated.

Rank order of the HI values are preserved (i.e. zero rank errors) in five dimensions (Fig. 2) while rank orders in dimension four (Fig. 3) are not preserved (144 rank errors). This determination of the dimension shape space is in accord with Perelson's and Oster's earlier estimate of a low dimension (Perelson *et al.*, 1979) resulting from qualitative arguments concerning the binding of B cells to random antigens.

Next, we verify that (a) the resulting inferred geometry of shape space is independent of the initial starting values for the coordinates used to minimize eqn (1), and that (b) the resulting relative positions of antigens and antisera are so highly constrained by the rank relations that grossly different geometries do not result. Consider multiple runs in five dimensions, with different initial starting values for the coordinates and possibly different final coordinates. Scale, translation, rotation, and reflection transformations are not of interest. To factor out these inessential transformations we evaluate the correlation

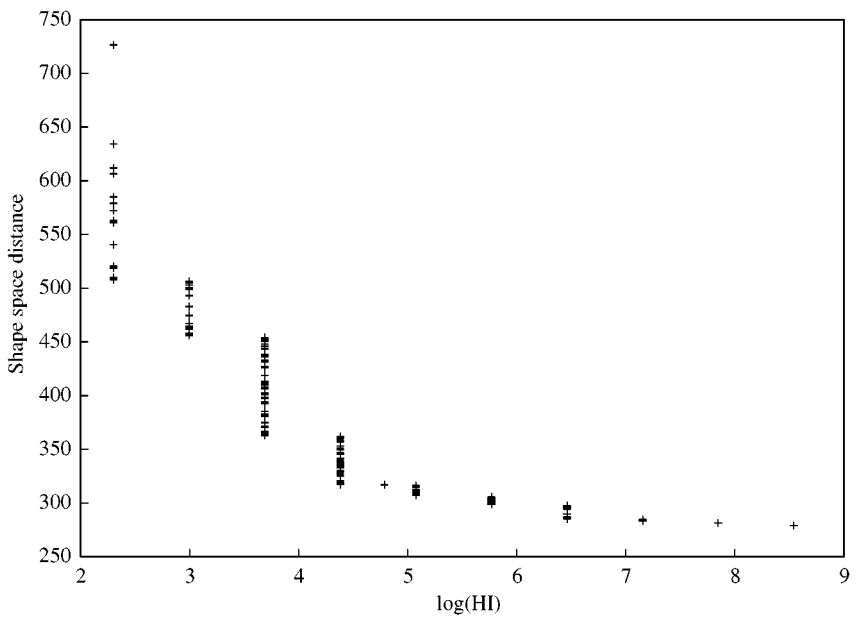

FIG. 2. Plot of log(HI) value on x-axis vs. computed distance in shape space (dimension five) on *y*-axis. Note that the rank order of distances in shape space agrees with the rank order of the experimental HI values in shape space dimension five.

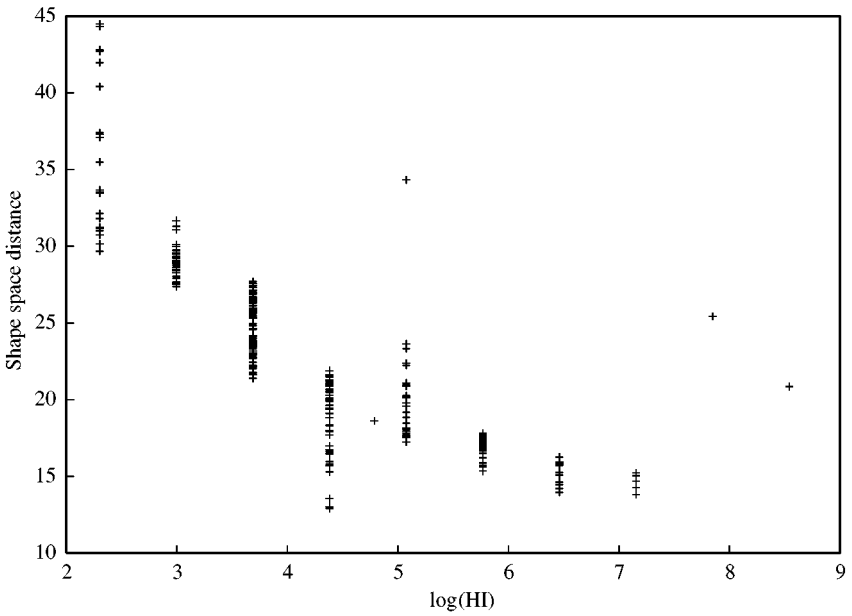

FIG. 3. Plot of log(HI) value on x-axis vs. computed distance in shape space (dimension four) on y-axis. Note that the rank order of distances in shape space does not agree with the rank order of the experimental HI values if the dimension of shape space is too low (c.f. Fig. 2).

(between pairs of runs) of the set of interpoint distances resulting from each run. A low correlation indicates variable geometries from run to run. Five initial choices of starting coordinates were used, resulting in $5*4/2 = 10$ possible pairwise correlation measures. The average of these ten correlations was 0.92, indicating a good reconstruction of the same underlying geometry from run to run. We conclude that the final computed geometry does not sensitively depend on the initial values of coordinates, that local minima are not a problem, and that the set of rank relations highly constrains the relative locations of points in shape space.

3.2. ANALYSIS OF HI ASSAY DATA: INFLUENZA H3 HEMAGGLUTININ 1968–1980

This data set comprises a panel of $M = 14$ antigens vs. $N = 14$ antisera published in Both *et al.* (1983) for strains selected from the years 1968–1980. The application of eqn (1) shows that rank order relations between experimental HI values and distances in shape space is preserved in dimension five, but not in lower dimensions (data not shown). We again conclude, using different experimental data, that the dimension of immunological shape space is low, on the order of five dimensions.

An evaluation of the correlation of interpoint distances between five different runs, analogous to the correlation analysis performed on the previous data set, results in an average interpoint distance correlation of 0.8. from run to run. The reason for this increased variability relative to the Raymond data set is that there happen to be antigens with low cross-reactivity (i.e. low HI values, and therefore associated high distance) to a large number of the antisera used in the data set (data not shown). Hence, the associated points in shape space are less constrained by the given data, which is reflected by a variability in the final computed distances. In the Raymond data set, as well as other data sets considered below, this issue, which is one of poor data for some points, is not a problem.

Of interest in this data set are the relative relationships of the strains A/HK/68, A/Eng/72, A/PC/73 and A/Vic/75 in shape space. These strains appear in the data collected from outbreaks of H3N2 influenza at Christ's Hospital in 1974 and in 1976. Smith *et al.*, (1999), suggest that patients vaccinated in successive years can expect to have higher attack rates when exposed to

infectious virus compared to first time vaccinees if the vaccine 1 to vaccine 2 distance is small, and the vaccine 1 to epidemic strain distance is comparatively medium or large. In 1974, the epidemic strain was A/PC/73-like and patients were previously vaccinated in successive years with the vaccine strains A/HK/68 followed by A/Eng/72. In 1976, the epidemic strain was A/Vic/75-like and patients were previously vaccinated in successive years with the vaccine strains A/Eng/72 followed by A/PC/73.

To visually represent the relationship of points in five-dimensional shape space we borrow a device from phylogenetic tree analysis, which visually represents relations between sequences given distances between them. The neighbor-joining algorithm is a classic tree building algorithm (Hillis *et al*., 1990) that uses a matrix of pairwise distance relations as input. Distances were calculated from the points assigned to the antigens and the antisera in the minimal five-dimensional space which preserved the rank order relationships of the HI values. Figure 4 is a neighbor-joining tree produced by the Phylip package (Felsenstein, 1993), illustrating the relation in five-dimensional shape space between the points of interest, including both antisera and antigen.

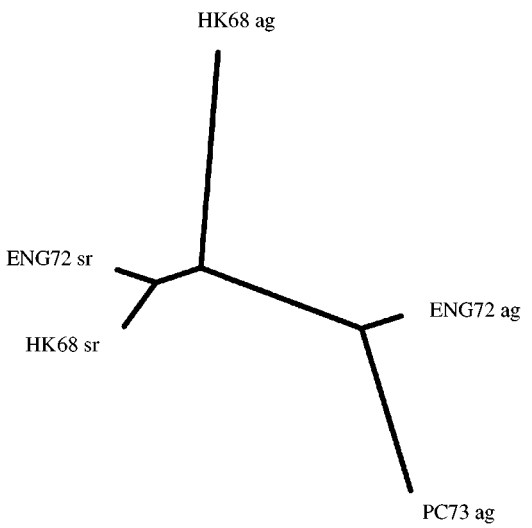

FIG. 4. A neighbor-joining tree, produced by the Phylip package, illustrating the relations between antisera and antigens in five-dimensional shape space. The names ending with "ag" denote antigens, those ending in "sr" denote antisera. Affinity or HI value is monotonically related to the distance between antigens and antisera: the smaller the distance, the higher the HI value. Antigen–antigen or antisera–antisera similarities are related to the respective antigen–antigen or antisera–antisera distances: the smaller the distance, the higher the similarity.

The vaccine 1 (HK68ag) to vaccine 2 (ENG72ag) distance is seen to be less than the vaccine 1 to epidemic strain (PC73ag) distance. The attack rate for first time vaccinees was 3% and for two time vaccinees was 11%, in accord with the suggestion by Smith *et al*. (1999). Similarly, for the 1976 outbreak the vaccine 1 to vaccine 2 distance is less than the vaccine 1 to epidemic strain distance. The attack rate for first time vaccinees was 13% and for two time vaccinees was 22%, also in accord with the suggestion.

3.2.1. *Repeated HI Tables*

Experimental variability can result in different reported HI values for the same set of antigens and antisera if the experiments are repeated on different days. We consider the effects on the computed geometry of this experimental variability. Data kindly provided by the Influenza Branch of the Centers for Disease Control and Prevention (priv. comm.) report results of five different HI assay experiments performed on the same set of antisera/antigens over five separate days in 1990: 8/10/90, 8/30/90, 9/26/90, 9/27/90, 10/2/90. The data comprised 11 antigens and 11 associated antisera for the following H3N2 influenza antigens, spanning a period from 1987 to 1990: BEIJING/337/89, BEIJING/353/89, CZECHOSLOVAKIA19/89, ENGLAND/ 427/ 88, ENGLAND/648/89, GUIZHOU/54/89, SHANGHAI/06/90, SHANGHAI/11/87, SHANGHAI/16/89, SICHUAN/68/89, VICTORIA/5/89.

In agreement with the analysis of other data analysed in previous sections, the five data sets can be represented without rank errors in a shape space of low dimension (dimension is five for the 8/10/90 data set, and dimension is four for the remaining data sets). Five different geometries resulted, yielding five sets of interpoint distances. Similar to the previous analysis, irrelevant scale, rotation, translation and reflection variations can be factored out by examining the correlations between sets of interpoint distances in the five different geometries. There are $5*4/2$ or ten such possible correlations between the five geometries

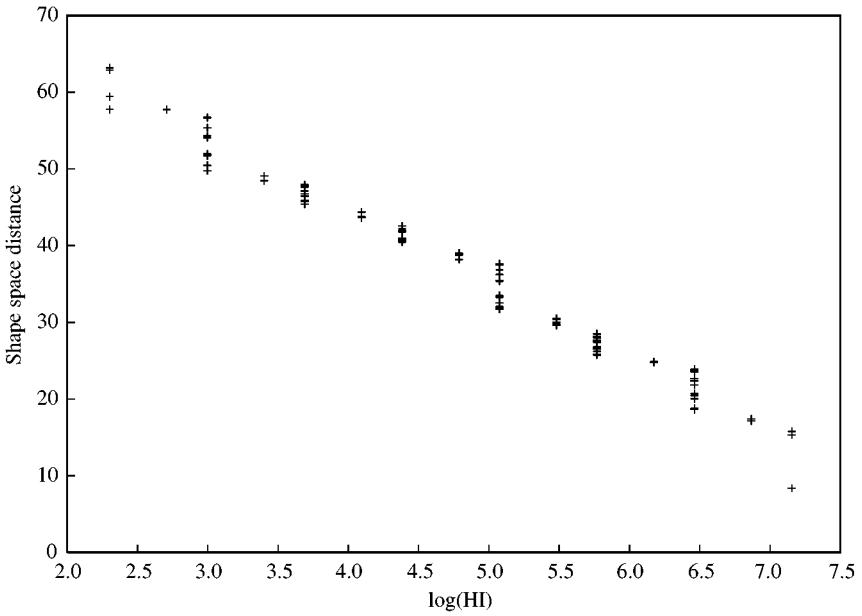

FIG. 5. Plot of log(HI) on x-axis vs. computed distance in shape space (dimension five) on y-axis. Note the linear relationship between log(HI) and shape space distance.

of the five data sets. The average of these ten correlations was 0.96, indicating a very good correspondence between the computed geometries, in spite of variation in the reported HI values used as input. Furthermore, plotting HI values vs. distance for this data shows that HI and distance are simply related by $HI = A \exp -(Distance)$, where A is a scale factor (see Fig. 5). Hence, ordinal MDS algorithm can be used to determine the monotonic transformation relating shape space distance to HI values, which may then be followed by metric MDS algorithms if desired.

3.2.2. *Validation/Simulation Studies*

HI measurements involve a total titration by a factor of approximately 1000 ($2^{10} = 1024$, corresponding to ten two-fold dilutions) which results in only ten possible distinct values appearing in any given HI panel. We address in simulation Study 1 (below) whether this relatively small number of discrete values could result in the algorithmic determination of an artificially low dimension even for high dimensional, random data, if such data are similarly binned.

3.2.3. *Simulation Study* 1

Artificial data in 15 dimensions was created by generating 14 points for antisera, and 19 additional points for antigens, with coordinates drawn from a Gaussian distribution with zero mean and unit variance in 15-dimensional space. To relate these generated distances to discrete, two-fold, HI values we first scale the distances between 0 and 1 for convenience, and then bin the distances to yield associated HI values as follows:

$$0.9 < Distance <= 1.0 \text{ implies } HI = 10,$$

$$0.8 < Distance <= 0.9 \text{ implies } HI = 20,$$

$$0.7 < Distance <= 0.8 \text{ implies } HI = 40,$$

...etc. This binning generates a set of simulated HI values 10, 20, 40, 80, 160, 320, 640, 1260, 2560, 5120 which are related by powers of two.

Next we determine if the data, generated in 15 dimensions, and filtered through the simulated two-fold HI dilution study (above) can be fit in a low dimensional, e.g. five-dimensional space.

3.2.3.1. *Result* 1. Five separate sets of binned data were generated as above. Fits in 15 dimensions were successful, as expected. The same data could not be fit in five dimensions. Similarly, five sets of data prepared in ten dimensions were successfully fit in ten dimensions, but not in dimension five. Finally, five sets of data created in seven dimensions was successfully fit in seven

dimensions, but not in dimension five. Hence, binned data comprising only ten discrete values, will not result in an algorithmically determined dimension that is significantly less than the real dimension of the space in which the points were generated. The determination of a dimension of approximately five for the real, experimental HI values would therefore seem significant.

3.2.4. *Simulation Study* 2

HI values of the experimentally determined data were randomly permuted within a given HI table to see if even such permuted data could be fit in low dimension.

3.2.4.1. *Result* 2. For conciseness, we only report results on the Raymond data set (Raymond *et al*., 1986) considered earlier. Five different attempts were made to fit the permuted data in five dimensions, using five different sets of initial values for the coordinates. All five attempts to fit the permuted data in dimension five were failures. Similarly, we attempted to fit the same permuted data in dimension ten, and also in dimension 15. In dimension ten, three of the five initial sets of random values for the coordinates resulted in failure to fit in ten dimensions. In 15 dimensions, all five sets of random initial values for the coordinates for Russian strain resulted in successful fits. It is not surprising that permuted data can be fit in high (ten or 15)-dimensional space, because in a high-dimensional space there is sufficient "room" to adjust the coordinates of only $14 + 19 = 33$ points to accommodate the small number of discrete HI values. Reassuringly, in a low dimensional space, e.g. five dimensions, there is not sufficient freedom.

4. Discussion

The computational techniques developed here, when applied to sets of experimental HI data for influenza, yield a consistent estimate of four to five dimensions for the dimension of immunological shape space. These techniques can:

(1) Deduce the dimension of shape space from experimental data and assign coordinates to both antisera and antigen in the shape space.

(2) Accommodate arbitrary monotone relationships between distance in shape space and experimental measurements related to affinity.

(3) Calculate antibody–antibody and antigen–antigen similarities based on experimental data quantifying antibody–antigen interactions.

(4) Accommodate imprecise data whose only significance may lie in the rank order of the experimental values.

After this work was completed we became aware of the work of B-Rao & Stewart (1996), which used metric MDS to reduce what we referred to in Section 2 as "panel distance" relations, to smaller dimensions; and the work of Beyer & Masurel (1985) and Weijers *et al*. (1985) in which panel distance relations were represented using "phylogenetic trees", similar to our tree representation in Section 3 of true shape space distance. The non-metric MDS approach presented here, which avoids use of "panel distance" with its associated problems (see Section 2), can infer a detailed geometry of immunological shape space given experimental data of limited precision such as a panel of hemagglutination inhibition assay data, or other measures of affinity. This ability is potentially of value in a number of application areas, such as analysis of HI data as part of the selection process for deciding components of the annual influenza vaccine.

Each HI table produces a separate shape space in which the antigens and antisera for that table are located. Since reference panels for successive years typically contain points which overlap, it is possible in principle to construct one large shape space (and resulting HI table) incorporating the results of several separate but overlapping assays. "Overlap" in this context means that the separate HI tables, e.g. reference panels used in successive years, contain some antigens and antisera in common. Hence, each shape space geometry will have a subset of points that have identical geometries to a subset of points in the shape space produced from another (overlapping) panel. Computationally aligning these overlapping subsets of points using a rigid body transformation then relates all the points of the different shape space geometries. The accuracy of the resulting global shape space geometry

(and equivalent global HI table) will depend on error propagation as successive geometries/tables are joined. In principle, however, such an approach could be used to describe the global evolution of antigenicity in shape space across decades of viral evolution. Phylogenetic trees are a standard way of representing viral evolution based on sequence data, and it will be of interest to relate shape space evolution as defined here (e.g. Fig. 4), to sequence space evolution.

An issue not addressed in this paper is the physico-chemical interpretation of the reconstructed coordinates of antigens and antibodies in the low-dimensional shape space. The low dimension (approximately, dimension five) of immunological shape space deduced herein for influenza from the analysis of HI data, indicates that approximately five independent combinations of molecular descriptors including e.g. geometric complementarity, electrostatic interactions, hydrophobicity, etc., serve to describe the binding of antibody to antigen. The work of Katchalski-Katzir *et al.* (1992) and Aflalo *et al.* (1994), and later work of (Palma *et al.* (2000), as well as related work in the cited references, is of interest in this regard. In this line of investigation, molecular descriptors are discovered which allow the prediction of docking configurations of proteins and ligands. How to relate these descriptors, or combinations of these descriptors, to the underlying low-dimensional shape space remains an important challenge.

The formalism presented here is independent of the specific application to influenza. It may be applied to other serological data, as well as to other affinity studies quantifying the binding of arbitrary molecules and ligands. Additional applications will be considered elsewhere.

This research was supported by the Department of Energy under contract W-7405-ENG-36. The authors gratefully acknowledge the hospitality of the Santa Fe Institute where part of this research was performed. Lapedes thanks Stuart Kauffman for helpful conversations in the early stages of this work, and Derek Smith for sharing prepublication results as well as for useful discussions in the latter stages of this work. The authors thank Nancy Cox and members of the Influenza Branch, Centers for Disease Control and Prevention, for generous provision of hemagglutination inhibition assay data for influenza and for useful feedback.

REFERENCES

AFLALO, VAKSER, I. & AFLALO, C. (1994). Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins* **20,** 320–329.

BEYER, W. & MASUREL, N. (1985). Antigenic heterogeneity among influenza A(H3N2) field isolates during an outbreak in 1982/83, eestimated by methods of numerical taxonomy. *J. Hyg. Camb.* **94,** 97–109.

BORG, I. & GROENEN, P. (1997). *Modern Multidimensional Scaling*. Springer Series in Statistics. Berlin: Springer Press.

BOTH, G., SLEIGH, M., COX, N. & KENDAL, A. (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *J. Virol.* **48,** 52–60.

BRAUN, W. (1987). Distance Geometry and related methods for protein structure determination from NMR Data. *Quart. Rev. Biophy.* **19,** 115–157.

CDC, private communication.

DEBOER, R., HOGEWEG, P. & PERESLON, A. (1992). Growth and recruitment in the immune network. In: *Theoretical and Experimental Insights Into Immunology*. NATO ASI Series, Vol. **H66**. Berlin: Springer-Verlag.

DEBOER, R., SEGEL, L. & PERESLON, A. (1992). Pattern formation in one and two dimensional shape space models of the immune system. *J. theor. Biol.* **155,** 295–333.

EDELMAN, S. (1995). Representation of similarity in 3D object discrimination. *Neural Comput.* **7,** 407–422.

EDELSTEIN, L. & ROSEN, R. (1978). Enzyme–substrate recognition. *J. theor. Biol.* **73,** 181–204.

FELSENSTEIN, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

HILLIS, D. & MORITZ, C. (eds) (1990). *Molecular Systematics*, Sunderland, MA: Sinauer Associates Inc.

KATCHALSKI-KATZIR, E., SHARIV, I., EISENSTEIN, M., FRIESEM, A., AFLALO, C. & VAKSER, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci.* U.S.A. **89,** 2195–2199.

KRUSKAL, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29,** 1–27.

KRUSKAL, J. (1964). Nonmetric multidimensional scaling. *Psychometrika* **29,** 115–129.

LEHMANN, E. (1975). *Nonparametrics*: *Statistical Methods Based on Ranks*. San Fransisco: Holden-Day.

PALMA, P., KRIPPAHL, L., WAMPLER, J. & MOURA, J. (2000). Bigger: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **39,** 372–384.

PERELSON, A. & OSTER, G. (1979). Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. theor. Biol.* **81,** 645–667.

PERELSON, A. (1988). Towards a realistic model of the immune system. In: *Theoretical Immunology*, Part Two. *SFI Studies in the Sciences of Complexity*. Reading, MA: Addison-Wesley.

B-RAO, C. & STEWART, J. (1996). Inverse analysis of empirical matrices of idiotypic network interactions. *Bull. Math. Biol.* **58,** 1123–1153.

RAYMOND, F., CATON, A., COX, N., KENDAL, A. & BROWNLEE, G. (1986). The antigenicity and evolution of influenza H1 haemagglutinin from 1950–1957 and 1977–1983: two pathways from one gene. *J. Virol.* **148,** 275–287.

SEGEL, L. & PERESLON, A. (1988). Computations in shape space: a new approach to immune network theory. in: *Theoretical Immunology*, Part Two. *SFI Studies in the Sciences of Complexity*. Reading, MA: Addison-Wesley.

SHEPHERD, R. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human Factors* **5,** 33–48.

SHEPHERD, R. (1964). Attention and the metric structure of the stimulus space. *J. Math. Psychol.* **1,** 54–87.

SMITH, D., FORREST, S. ACKELY, D. & PERELSON, A. (1999). Variable efficacy of repeated annual vaccinations against influenza: an in-machina study. *Proc. Natl Acad. Sci. U.S.A.* **96,** 14 001–14 006.

WEIJERS, T. OSTERHAUS, A., BEYER, W., VAN ASTEN, J., DE RONDE-VERLOOP, R, BIJLSMA, K. & DE JONG, J. (1985). Analysis of antigenic relationships among influenza virus strains using a taxonomic cluster procedure: comparison of three kinds of antibody preparations. *J. Virol. Meth.* **10,** 241–250.