

# The Geometry of Shape Space: Application to Influenza

Alan Lapedes  
Robert Farber

*Theoretical Division, MS B213  
Los Alamos National Laboratory  
Los Alamos, NM 87545  
and The Santa Fe Institute  
1399 Hyde Park Road  
Santa Fe, NM 87501*

Address correspondence and reprint requests to Alan Lapedes, MS B213, Los Alamos National Laboratory, Los Alamos NM, 87545. E-mail [asl@lanl.gov](mailto:asl@lanl.gov), phone: 505-667-7608, fax: 505-665-3003

Running head: Geometry of Shape Space

Keywords: influenza, hemagglutination assay, shape space, affinity, multi-dimensional scaling, antibody, antigen

## Abstract

Shape space was proposed by Perelson and Oster 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 thought of as 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. Up to now, coordinates of points in shape space have been purely implicit.

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 assays, or other general affinity assays. An application of these algorithms to panels of hemagglutination inhibition assays for influenza show that the dimension of immunological shape space is low, roughly dimension five, in accord with Oster’s and Perelson’s earlier qualitative estimates.

Explicit numerical values are provided by the algorithms for co-ordinates of molecules in shape space, whereas previously such coordinates have been a conceptual construct and totally implicit. 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 analysis of antibody-antigen interactions and are generally applicable to affinity data resulting from binding assays.

## 1 Introduction

“Shape space” was introduced by Perelson and Oster [Perelson (1979)] in 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 et. al. (1988)] [Perelson (1988)] [DeBoer et. al. (1992)] [DeBoer et. al. (1992)]. This paper presents algorithms related to ordinal and metric multidimensional scaling [Shepherd(1963)], [Shepherd(1964)], [Kruskal(1964a)], [Kruskal(1964b)] 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 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 (1979)] [Segel et. al. (1988)] [Perelson (1988)] [DeBoer et. al. (1992)] [DeBoer et. al. (1992)] 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 (ap-

proximately 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 explicit coordinates for molecules in shape space given experimental data, such as hemagglutination inhibition (HI) assays. 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

**The Computational Problem:** It is assumed that a panel of affinity related measurements, such as hemagglutination inhibition assays, are available for a set of  $M$  antigens reacting with  $N$  antisera comprised of antibodies.

**The full computational problem is** to use the experimental data to deduce coordinate vectors in shape space for both the antigens and the antibodies i.e. to simultaneously construct  $N + M$  coordinate vectors describing  $N + M$  points in the same space. Distances from antigens to antibodies, as calculated from the coordinate vectors, are to be monotonically related to their measured HI values, while distances from antigen to antigens, respectively antibodies to antibodies, define inferred similarities among these points.

**A subset of the full computational problem is** the definition of coordinates for either the set of antigens, or for the set of antibodies, but not both simultaneously. Similarity is then defined as the distance between points in this space. Affinity can not be represented. We first point out difficulties in the conventional approach to this sub-problem, and then present solutions to the full problem.

**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 versus  $N$  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. These vectors de-

finds coordinates in a space whose dimension is the (arbitrary) number of antibodies, respectively, antigens, in the panel. The distance, or (dis)similarity between antigens, respectively antibodies, can then be calculated in the usual Euclidean fashion from these coordinates by forming the sum of the squares of coordinate differences. This procedure is often used (at least implicitly) to examine hemagglutination inhibition assay data and to infer similarities between antigens, or between antibodies, based on experimental data.

Clearly, defining simultaneous coordinates for both antigens and antibodies (i.e. placing both antibodies and antigens in the same space) is not possible using the simple procedure above, since by construction, coordinates are only provided for either antigens or antibodies, but not both. Furthermore, this simple method to define similarities can be highly inaccurate when affinities are 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 five dimensions. Hence  $N$  of a simulated  $NM$  panel is  $N = 10$ . Next, an additional fifty points is scattered in the space in a similar fashion to represent fifty antigens in the five dimensional shape space. Hence,  $M = 50$ . The fact that experimental panels do not usually have e.g. fifty antigens is irrelevant. We use fifty antigens in this  $50 * 10$  example merely to make trends visually apparent in Fig. (1).

To illustrate the problem with simplistic approaches to similarity definition we concentrate on the set of distances within the set of antigens (one could equally well consider the set of antibodies), as defined above. We also consider the distances within the set of fifty antigens as calculated in the true underlying five dimensional shape space, and compare this to the associated set of distances as calculated from coordinates defined by elements of the rows of the  $50 * 10$  panel (see above). This simplistic method represents the fifty antigens of this example as points in a ten dimensional space. However, the true underlying dimension of the space in this example is five dimensional. The effect of this mismatch can be seen in Fig (1). The true five dimensional shape space distance is plotted on the x-axis, while the ten dimensional distance computed in the fashion described above from the  $50 * 10$  panel, is plotted on the y-axis.

It may be seen that attempting to select similar points based on the ten dimensional “panel distance”, by horizontally slicing in y-value, results in a wide range of associated distances in the true five dimensional space represented by x-value. Hence, for all but the very smallest y-values, the simplistic procedure to define similarities results in a wide range of similarity values in the true space, including significantly large x-values, i.e., quite poor similarities in the true space.

A more serious problem with “panel based methods” is that they can not represent antigens and antisera in the same space. Also the very dimension of the space of antigens or the space of antisera depends on the arbitrary panel size. Algorithms are required which can recover the true underlying shape space, which situate both

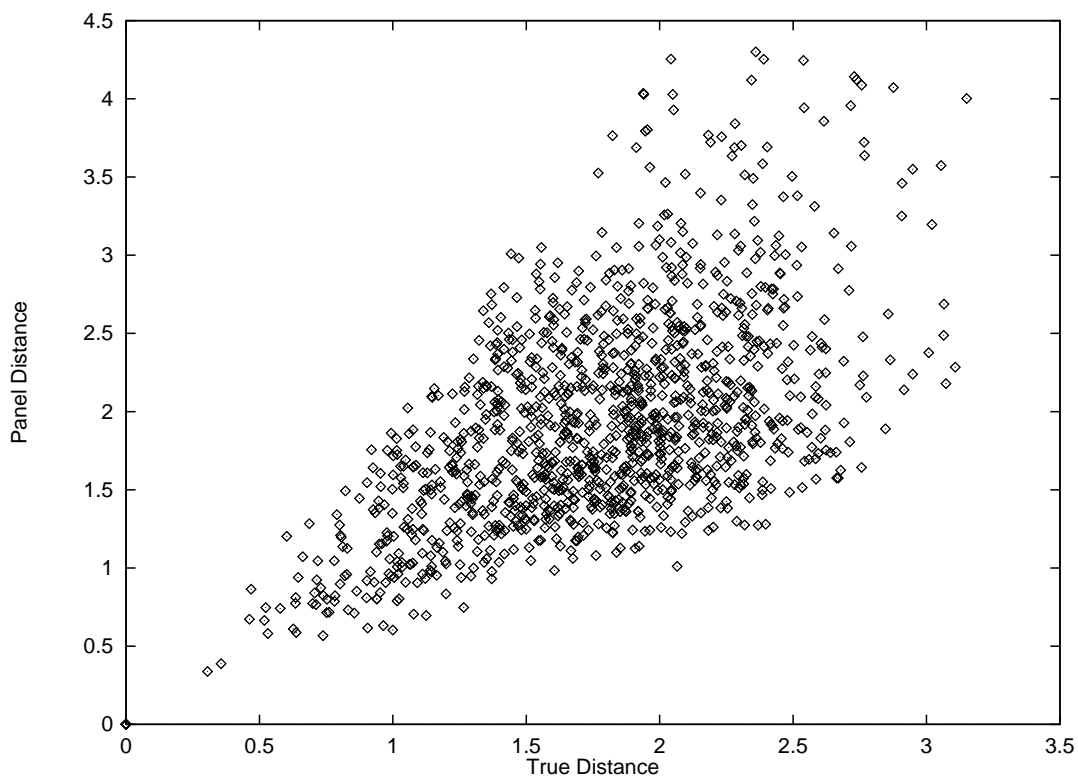


Figure 1: Scatter plot for antigens with true shape space distance on the x-axis, versus panel distance on the y-axis, for a panel of 50 antigens and 10 antibodies in a five dimensional shape space. Small distance corresponds to high similarity. Horizontal slices in y-value (i.e. fixed “panel distance”) 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 accurate similarity determination.

antigens and antibodies in the same space so that affinities as well as similarities are represented, and which defines a unique dimension of shape space. The algorithms presented here, related to techniques of multidimensional scaling, achieves these goals.

### **Similarity Measures From Affinity Panels: A New Approach**

“Multidimensional scaling” algorithms (referred to hereafter as MDS) are a class of algorithms initially developed in the computational psychology literature [Shepherd(1963)], [Shepherd(1964)] which reconstruct the relative coordinates of points, given only (possibly transformed) 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

A related class of algorithms called “distance geometry algorithms” has seen application in biology and chemistry in a quite different problem area than that considered here, i.e., the deduction of three dimensional coordinate information given nuclear magnetic resonance (NMR) data information about distances between atoms in molecules [Braun (1987)]. The shape space problem considered here differs in at least four fundamental ways from distance-geometry problems restricted to three dimensions:

(1) the dimension of shape space is not known and must be inferred from the experimental data

(2) the experimental data does not directly record distances, instead the experimental data (i.e. affinity measurements) are assumed to be related by an unknown monotonic transformation to distances in shape space

(3) data is available concerning only a subset of the distances, because affinity measurements relate antigen to antibody, and do not directly relate antigen to antigen, or antibody to antibody. In other words, some of the distances are “missing”. An important feature of our analysis is that such missing data, relating to measures of antigen-antigen similarity and to antibody-antibody similarity, can be calculated after application of our algorithms to measurements of antibody-antigen interaction.

(4) the experimental data is typically of low precision for which only rank order relationships may have significance.

An extension of multidimensional scaling to situations where only an ordered list of the *ranks* of the distances, and not the actual distances themselves are given, is called “ordinal multidimensional scaling” [Kruskal(1964a)], [Kruskal(1964b)] and has 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 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. 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. It is not uncommon in MDS ap-

plications to achieve highly accurate, quantitative reconstruction of coordinate values given only rank order information about distances between points, as may be verified by simulation (data not shown, see also [Edelman(1995)]) . The lower the dimension of the space, the less data is 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 (1979)] we assume that the affinity measurement for interactions between antibodies and antigens are monotonically related to the distances between the points describing them in shape space. High experimental affinities translates to small distance in shape space. The form of this monotonic relation is *a priori* unknown, but can be recovered via application of the algorithms presented here. In our context, 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. Alternatively, one could introduce a parameterized monotonic transformation into the MDS algorithm which relates distances calculated in shape space to experimentally measured affinities (an approach not pursued here). We note that the shape space problem is more general than problems usually considered in simple MDS analyses. For shape space, one is given distances or more generally experimental measurements monotonically related to distances, between only a subset of the data: only the antigen-antibody “distances” are given, and the antibody-antibody and antigen-antigen distances are to be inferred. This is known as the “unfolding problem” in the ordinal MDS literature [Borg(1997)].

**Algorithms:**

Various flavors of multidimensional scaling algorithms exist [Borg(1997)]. A unifying viewpoint is that such algorithms seek to minimize an objective function which is a function of (a) experimental data, assumed to be monotonically related to the distances between points, and (b) the distances between points, as computed from (to be determined) coordinate vectors representing each point in a space of assumed dimensionality. In our case, given experimental measurements representing the affinity of a panel of antigens and antibodies, the computational task is to determine the dimension of the space, and the co-ordinate vectors in that space representing each molecule, such that the computed distances between antibodies and antigens in the space are monotonically related to the experimental data.

**Metric MDS** For pedagogical purposes we first assume that the experimental measurements are in fact distances, and not more generally, monotonic transformations of distance. Versions of ordinal MDS will later be used to address experimental values which are monotonic transformations of distance. If the experimental data is assumed to be direct measurements of distances in the space (an assumption removed later) then the computational task of MDS may be formulated as minimizing the following objective function

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

where  $D_{ij}^{expt}$  are the known, experimentally determined, “distances” (or more generally, affinity measurements), and the estimated Euclidean distances  $D_{ij}$ , are computed in standard fashion given the coordinate vectors,  $X$ , of the points in a space of dimensionality  $D$  ( $D$  to be determined):

$$D_{ij} = |X_i - X_j|^2$$

Here  $||^2$  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

$$X_i = (X_i^1 X_i^2 X_i^3 \dots X_i^D)$$

where the numerical values of the components, and the dimension  $D$ , are to be determined.  $E$  will have the value zero if distances between points in shape space,  $D_{ij}$  match the observed distances,  $D_{ij}^{expt}$ .

$E$  is a function of the distances,  $D_{ij}$ , which in turn are a function of the components of the vectors,  $X_i$ . Hence if a dimension,  $D$ , is assumed, then one may optimize  $E$  as a function of the components using, e.g., steepest descents or conjugate gradient methods, to find optimal coordinates. Although local minima can in principle be a problem, we have found that good minima of the objective functions used in this paper are generally reached using conjugate gradient algorithms, and that choosing different random initial conditions result in minima which yield similar final coordinates (modulo global rotation, reflection, and translation, which are not of interest).

Because  $N$  points can be embedded in  $N - 1$  dimensions (e.g. three points define a plane in two-dimensions) a non-trivial embedding is obtained only if  $D$  is substantially less than the number of points. One can test all possible embedding dimensions, starting with highest dimension taken to be one less than the number of points. Just how pronounced the “knee” is in a plot of final  $E$  versus  $D$  is a function of the quantity and quality of the experimental data.

We also note that since some distances are missing in the case of affinity data, achieving a good reconstruction relies on the assumption that the  $M$  antigen points and the  $N$  antibody points are intermingled. If instead, the “center of mass” of e.g. the antigens were significantly displaced from that of the antibodies (i.e. if in actuality they formed two distinct clouds of points, as opposed to two intermingled clouds in shape space) then knowing the distances between points of the two clouds is not as constraining as if the clouds were intermingled. Intuitively, this point is made clear by thinking of a two dimensional example with two clouds of points, either (1) intermingled/overlapping, or (2) significantly displaced and not overlapping.

Various weighting factors can be introduced in  $E$ , if desired [Kruskal(1964a)], [Borg(1997)]. For example

$$E = \sum_{ij} (D_{ij}^{expt} - D_{ij})^2 / D_{ij}^2$$

will weight the smaller distances more heavily, and therefore place importance on the higher affinity values. To more directly relate MDS algorithms to solutions of the shape space problem, where high affinities correspond to small distance in shape space, one can introduce parameterized monotonic transformations,  $f(D_{ij}^{expt})$  into the function  $E$

$$E = \sum_{ij} (f(D_{ij}^{expt}) - D_{ij})^2 / D_{ij}^2$$

such that large  $D_{ij}^{expt}$  corresponds to small distance. Examples of such functions are  $f(X) = C - X$ , where  $C$  is a fixed positive constant converting large values of  $D_{ij}^{expt}$  to small values of  $D_{ij}$ . Alternatively, one might try  $f(X) = C \exp(-X)$  or similar transformations. *A priori* the correct monotonic transformation between distances in shape space and experimental values of any given data set is unknown. It may be verified via simulation (data not shown) that if “experimental values” involve one function,  $f_1()$ , and if the attempted reconstruction of shape space involves a different function (used in the minimization of  $E$ , above),  $f_2()$ , then (a) the reconstructed dimension is typically artificially inflated, and (b) a spread of reconstructed distances against true distances (similar to Figure (1)) is obtained. Presumably, if an assumed parametric functional form with adjustable parameters is sufficiently general, e.g. involving monotonic splines, then such errors could be mitigated, however we do not pursue that approach here.

Ordinal MDS, considered next, avoids these problems. Ordinal MDS algorithms result in a completely general, nonparametric description. After analysis by ordinal MDS algorithms, a parametric description of the relation of HI value to distance in shape space may be determined by inspection. We find that the relation between HI values and distance in shape space (which in principle can be any monotonic function) is described by the following simple parametric form (see Results).

$$HI = C * e^{-(distance)}$$

where  $C$  is a positive constant. If desired, it is possible to return to a metric MDS analysis, using the parametric form (above) obtained by analysis of ordinal MDS results, in order to account for errors in the experimental determination of HI values via the Least Mean Squares formulation of metric MDS.

**Ordinal MDS** Ordinal MDS addresses the monotonic transformation problem by using only rank order information, in which only the relative magnitude of the affinities and of the distances between points in shape space matter. The numerical values of the experimental data are not used other than to sort the values in rank order. A surprising and quite valuable property of ordinal MDS style algorithms is that metric information (i.e. relatively accurate floating point values for points in the underlying true space) can often be recovered from nonmetric (i.e. rank) information. This occurs because imposition of the set of rank inequalities effectively divides the space into sufficiently small regions such that coordinates for the points are highly constrained.

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 such value is assumed to be monotonically related to a distance  $D_{ij}$  between antigen  $i$  and antisera  $j$ . Order the experimental values from high to low, and index the resulting list of  $E_{ij}$  and the associated  $D_{ij}$  with a number from 1 to  $MN$ , such that rank one is the highest experimental value which is to be associated with the smallest distance. An objective function, which when minimized as a function of shape space coordinates ranks the computed distances in shape space in the same order as the experimental values, is:

$$E = - \sum_{i=1}^{i=MN} \log(g(D_{i+1} - D_i)) \quad (1)$$

Here,  $D_i$  references the  $MN$  computed distances between pairs of points representing antigens and antibodies in shape space, using the index based on the rank ordering of the experimental values explained above. Implicit in the notation for  $D_i$  is the fact that each  $D_i$  is defined from a particular antigen/antisera pair of vectors from which the distance is being computed.  $g()$  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.

This objective function is minimized as a function of the antigen/antisera coordinates in shape space, and achieves value equal to zero, when the rank order of the computed distances agrees with the rank order of the experimental values. This saturates the  $g()$  function near the value 1.0 and sends the  $\log()$  terms toward zero. 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. For the data considered below, minimization of the above objective function has proven simpler and more robust than the technique of monotonic regression [Borg(1997)] which has been described in the literature to accomplish ordinal MDS using a different objective function (however extensive comparisons were not performed). Extensions of the above objective function are possible, including a “soft” version (using sigmoidal functions) of Kendall’s coefficient of rank order correlation,  $\tau$  [Lehmann(1975)].

Even though rank information (as opposed to quantitative values) can be used to define coordinates in shape space, experimental precision still remains an issue. If, for example, hemagglutination inhibition assay values which do not differ by a factor of four are essentially indistinguishable due to experimental precision, as is typically the case for standard influenza serological assays analyzed below, then ambiguity can occur in the rank ordering of such “tied” experimental values. Other assays may have different (hopefully better) precision. Analysis of repeated HI assays for a given panel of influenza antigens and associated antisera (data kindly provided by the Influenza Branch, Centers for Disease Control and Prevention [CDC]) shows that the underlying

geometry of immunological shape space is recovered to good precision in spite of such experimental uncertainties (see Results).

Note that the algorithms described here for computing the geometry of shape space provides *simultaneous* coordinates for both antigens and for antibodies. Hence antigen-antigen and antibody-antibody distances can also be calculated, even though experimental information is provided only for antigen-antibody interactions. Antigens and antibodies are represented by points in the same space where the distances between antigens and antibodies is, by construction, related to their affinities. On the other hand, the antigen to antigen distances and the antibody to antibody distances quantify the similarity among antigens, respectively antibodies, in regards to the interaction(s) being measured. Since visualization of the relative positions of points in a space higher than three dimensions is difficult, standard techniques such as neighbor-joining [Hillis(1990)] may be used to provide dendograms which illustrate the relations between points in dimensions higher than three. Investigation of the relation of such “HI dendograms” to phylogenetic trees [Hillis(1990)], produced by sequence analysis using the DNA sequences of the antigens, could potentially be used to illuminate the relation between sequence changes and associated antigenic changes.

### 3 Results

Influenza is a rapidly mutating RNA virus for which there is a national and international policy of annual vaccination. Abundant HI data is 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 is 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(1986)], [Both(1983)], five unpublished HI panels [CDC] 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 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 dimension equal to either four or five. .

**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) versus 14 antisera (hence,  $M = 19$  and  $N = 14$ ) for influenza H1 subtype hemagglutinin, published in an investigation [Raymond(1986)] of the evolution and antigenicity of the re-emergent influenza strain A/USSR/90/77. 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. A/USSR/90/77 influenza is of interest, in part, because of the re-emergence in 1977 of a strain (A/USSR/90/77) remarkably similar to one first observed in 1950.

Application of Eqn.(1) in successive dimensions results in the rank order of the hemagglutination inhibition (HI) assay values published in [Raymond(1986)] being preserved in a minimum of five dimensions. A clear signature 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  $D_{i+1} > D_i$  of Eqn. (1) is violated. This signature may be seen graphically in Figures (2) and (3). In Fig.(2) the log of the experimental HI values are plotted on the x-axis, versus the associated distance between antigen-antisera in five dimensions, on the y axis.

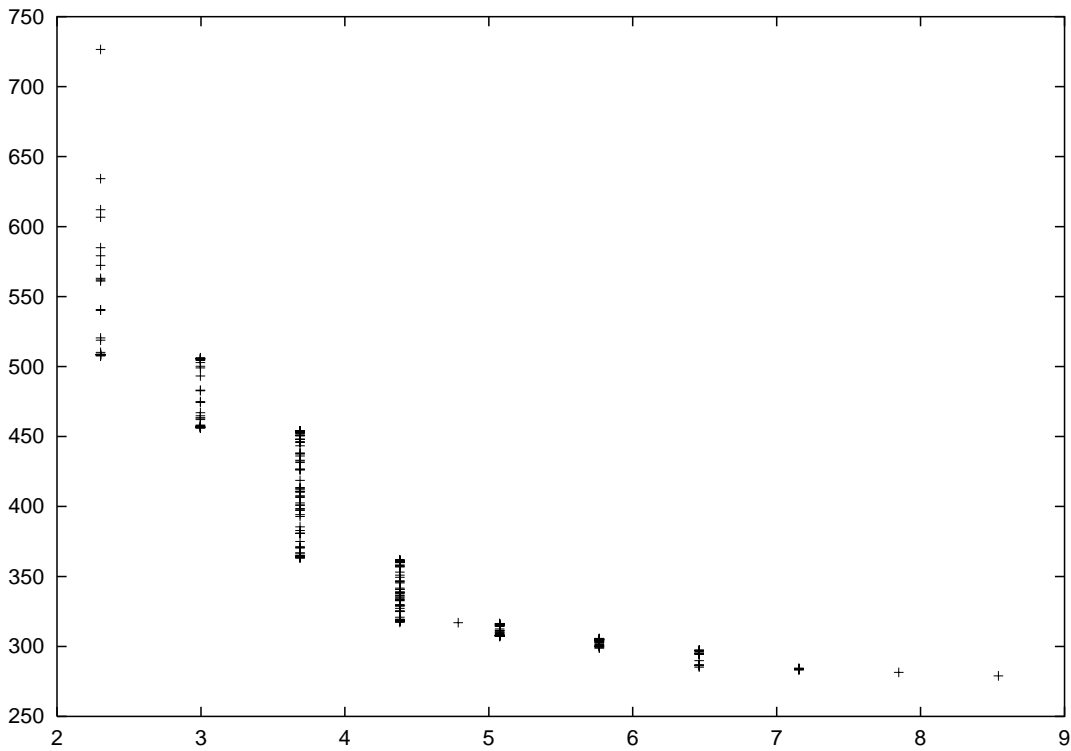


Figure 2: Plot of log(HI) value on x-axis versus computed distance in shape space (dimension five) on y-axis. Note that the rank order of distances in shape space agrees with the rank order of the experimental HI values in shape space dimension five.

Similarly, Fig.(3) shows the result of the same algorithm in four dimensions.

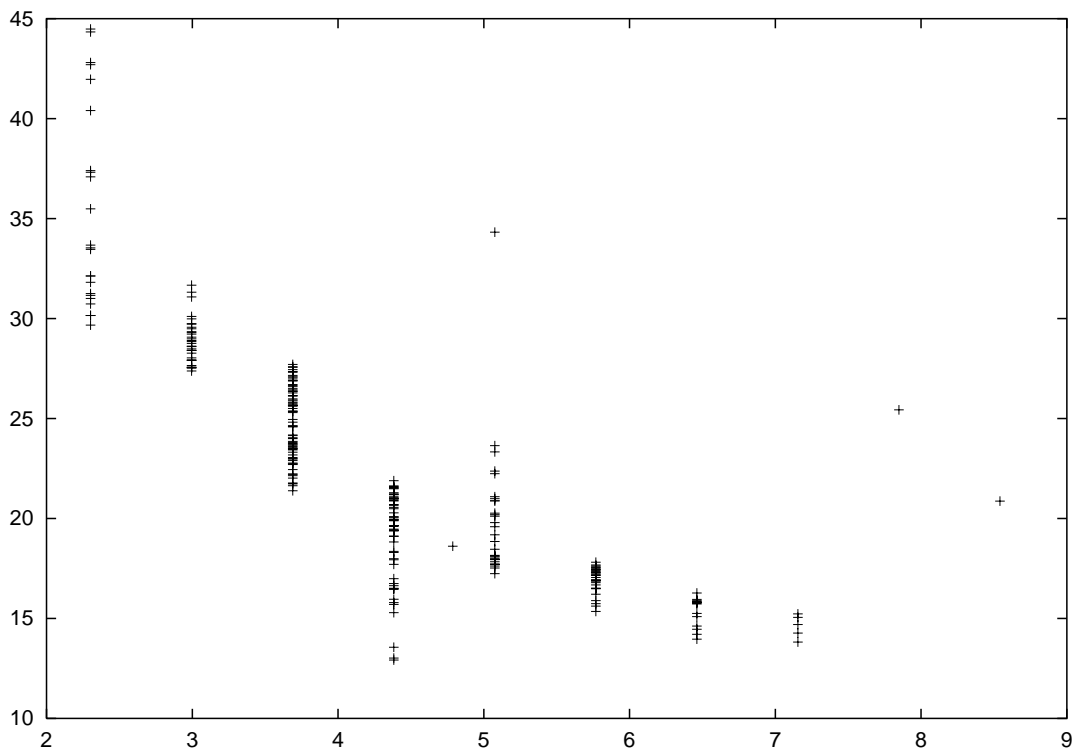


Figure 3: Plot of  $\log(\text{HI})$  value on x-axis versus computed distance in shape space (dimension four) on y-axis. Note that the rank order of distances in shape space does not agree with the rank order of the experimental HI values if the dimension of shape space is too low, e.g. four dimensional as opposed to five dimensional (*c.f.* Fig. (2)).

It may be seen that rank order of the HI values are preserved (zero rank errors) in five dimensions (Fig. (2)) while rank orders in dimension four (Fig.(3)) are not preserved (144 rank errors), thereby providing a numerical as well as graphical signature of the dimension of immunological shape space. This determination of the dimension shape space is in accord with Perelson's and Oster's earlier estimate of a low dimension [Perelson (1979)] resulting from qualitative arguments concerning binding of B-cells to random antigens.

It is important to 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 the rank relations can not be satisfied with grossly different geometries. To illustrate these points consider multiple runs in five dimensions, with different initial starting values for the coordinates. Different coordinate values which are related by scale, translation, rotation, and reflection transformations

can result. To separate such inessential transformations from possible variations in the geometry from run to run which are related either to either (a) local minima in Eqn. (1) and/or (b) to the imprecise nature of rank order constraints, we evaluated the correlation between pairs of runs of the set of interpoint distances resulting from each run. Such a correlation measure of interpoint distances is insensitive to inessential rotation, reflection, translation, and scale changes from run to run. Low correlation is an indicator of 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 10 correlations was 0.92, indicating a good reconstruction of the same underlying geometry from run to run. We conclude the final computed geometry does not sensitively depend on the initial values of coordinates used in the computation and that the set of rank relations highly constrains the relative locations of points in shape space.

### **Analysis of HI Assay Data: Influenza H3 Hemagglutinin 1968-1980**

The second data set we consider is HI assay data from a panel of  $M = 14$  antigens versus  $N = 14$  antisera published in [Both(1983)] for strains selected from the years 1968-1980. Application of Eqn.(1) in successive dimensions 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). A clear signature is therefore again obtained, using different experimental data, that the dimension of immunological shape space is low, on the order of five dimensions.

Evaluating the correlation of interpoint distances between five different runs, analogous to the correlation analysis performed on the previous data set, results in a average interpoint distance correlation from run to run of 0.8. Hence there is more variability in the geometry resulting from the Both data set than that of the Raymond data set. The reason for the lower correlation (more variability of the computed geometry from run to run) for the Both data set is because in this data set there are 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. This is reflected in the final computed distances, where small interpoint distances are highly correlated across the five runs, but large distances (associated with low HI values) are poorly correlated across runs, resulting in the average correlation of 0.8. This makes intuitive sense: if specific antigens (respectively, antisera) have low reactivity across the set of antisera (respectively, antigens) then the points in shape space corresponding to those members of the HI table with low reactivity will be poorly constrained. HI panels in which any given antigen (respectively, antisera) has a measurable cross-reactivity with a minimum number of antisera (respectively, antigens), solves this problem, which is an issue of the data and not one intrinsic to the algorithm. In the Raymond data set, as well as other data sets considered below, this issue is not a problem, as evidenced by the higher average correlation of interpoint distances from run to run.

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 data collected from outbreaks of H3N2 influenza at Christ's Hospital in 1974 and in 1976. Smith et. al., [Smith(1999)], suggest that patients inoculated in successive years can expect to have higher attack rates of 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 inoculated 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 inoculated 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 used in sequence analysis. Phylogenetic trees are a convenient method to represent the distance relationships between sequences in such a way that the relations can be viewed in a two-dimensional drawing. The neighbor-joining algorithm is a classic tree building algorithm [Hillis(1990)] that uses a matrix of pairwise distance relations as input. We use the same device to represent distance relationships in shape space. 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. Fig. (4) is a neighbor-joining tree produced by the Phylip package [Phylip(1993)] which illustrates the relation in five dimensional shape space between all points, including both antisera and antigen.

The vaccine 1 (HK68ag) to vaccine 2 (ENG72ag) distance is seen to be less than the 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(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.

Such trees displaying relationships in shape space, enable one to begin to see the antigenic evolution of influenza over time, similar to the way standard phylogenetic trees [Hillis(1990)] display sequence evolution over time. However, the available HI data is broken up into separate (and often overlapping) tables covering different time periods, and it will be necessary to link the different geometries arising from separate tables together into one global geometry to attempt to survey the antigenic evolution of influenza over broad time scales (see Discussion).

**Repeated HI Tables** Experimental variability can result in different reported HI values for the same set of antigens and antisera if experiments are repeatedly performed on different days. Here we consider the effects on the computed geometry of experimental variability in determining HI values. Data kindly provided by the CDC (private communication [CDC]) reports 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 anti-

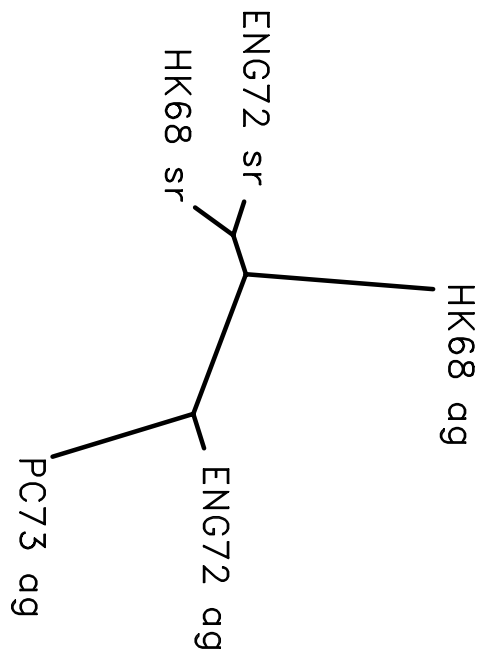


Figure 4: A neighbor-joining tree, produced by the Phylip package, illustrating the relations between antisera and antigens in five dimensional shape space. Names ending with “ag” denote antigens, those ending in “sr” denote antisera. Affinity or HI value is monotonically related to the distance between 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.

gens and 11 associated antisera for the following H3N2 influenza antigens, spanning a period from 1987 to 1990: BEIJING/337/89, BEIJING/353/89, CZECHOSLOVAKI19/89, 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 analyzed 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). Of interest is the variability in the computed geometry of the points across the five different data sets. Each of the five data set was subjected to an ordinal MDS analysis (see Material and Methods) in five dimensions. Five different geometries resulted, yielding five sets of interpoint distances.

Similar to the previous analysis which compared variation in geometries across different initial starting coordinates, irrelevant scale, rotation, translation and reflection variations can be factored out of the results of these five repeated data sets by examining the correlations between sets of interpoint distances in the five different geometries. There are  $5 * 4/2$  or 10 such possible correlations possible between the five geometries of the five data sets. The average of these 10 correlations was 0.96, indicating a very good correspondence between the computed geometries, in spite of variation in the reported HI values across the five data sets due to experimental precision.

A striking relation between distance in shape space and the HI value, corresponding to  $HI = e^{-(Distance)}$ , may be seen in Fig. (5). This illustrates a potential staged use of ordinal MDS algorithms followed by metric MDS algorithms. Ordinal MDS can be used first to deduce the *a priori* unknown monotonic relationship between HI value and shape space distance. Then metric MDS can employ this empirically determined relation to further improve the fit.

**Validation/Simulation Studies:** Further validation addressing issues of experimental precision and small sample sizes is warranted. In this section we report simulation results which address the issue that the experimental HI data results from two-fold dilution studies and hence the reported HI values are “binned” into a set of distinct values. The experimental protocol involves a total titration by a factor of approximately 1000 ( $2^{10} = 1024$ , corresponding to 10 two-fold dilutions) which results in only 10 possible distinct values appearing in any given HI panel. The first issue we address in simulation Study 1 (below) is 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 is similarly binned.

**Simulation Study 1** We create artificial data in a high dimensional space, e.g. 15 dimensions, 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. It is necessary to transform the distances between the antigens and the antisera to simulated HI values associated with two-fold dilution studies. HI values resulting from two-fold dilution studies can be assumed to take values in the discrete set 10, 20, 40, 80, 160, 320, 640, 1260, 2560, 5120. These numbers

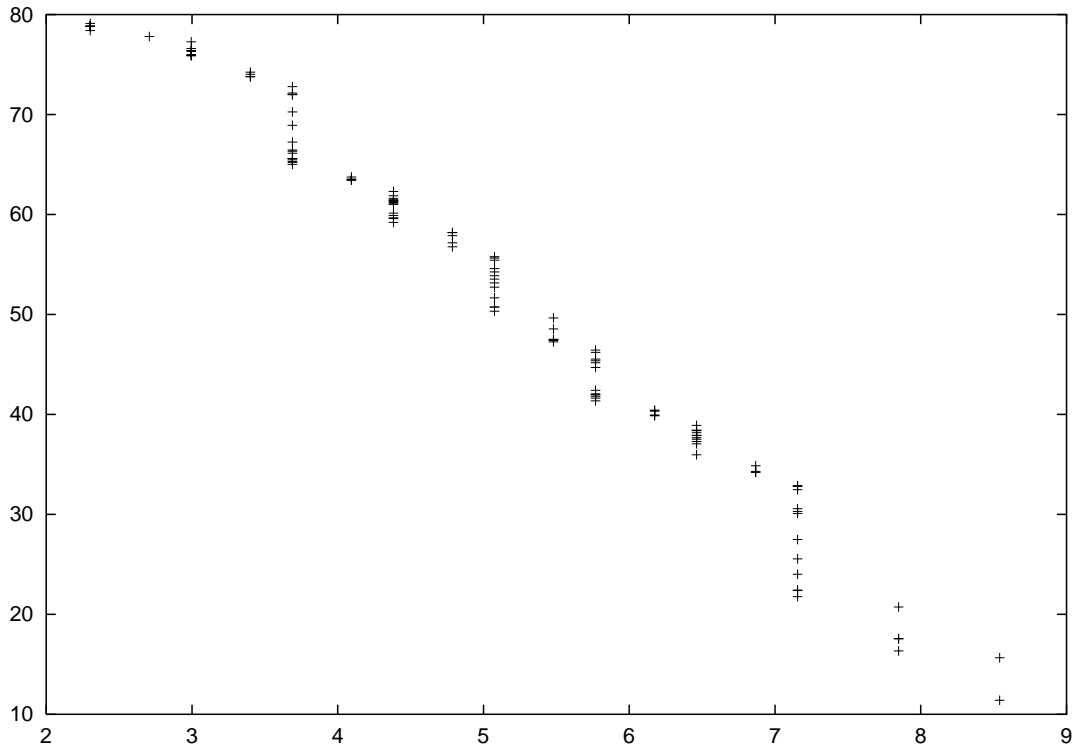


Figure 5: Plot of distances as computed in shape space on the y-axis for one of the repeated data sets, versus (the log) of the associated HI value on the x-axis. The linear relationship evident in the graph is a consequence of the following simple empirical relation between HI values and distances in shape space  $HI = e^{-(Distance)}$

are powers of two, multiplied by a scale factor of 10, which was introduced to produce simulated “HI tables” which have the appearance of experimentally determined tables. The factor of 10 is irrelevant to the numerical analysis and is introduced for aesthetic reasons.

To relate these values to the generated distances, we first scale the distances between 0 and 1 for convenience, and assign HI values to the distances as follows:

$0.9 < Distance \leq 1.0$  implies  $HI = 10$

$0.8 < Distance \leq 0.9$  implies  $HI = 20$

$0.7 < Distance \leq 0.8$  implies  $HI = 40$

...etc.

Next we determine if the data, generated in fifteen dimensions, and filtered through the simulated two-fold HI dilution study (above) can be fit in a low dimensional, e.g. five dimensional space. This addresses whether the dimension of shape space can be artificially estimated as being low, even when the generating points exist in high dimension, as a consequence of two fold dilution studies producing data comprised of a small number of discrete values which is fit by the algorithm.

### **Result 1**

Five separate sets of binned data, generated in fifteen dimensions as described above, were first fit in fifteen dimensions, and as one would expect the fit was successful. However, the same data generated in fifteen dimensions could not be fit in five dimensions. Similarly, five sets of data prepared in ten dimensions was 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.

All runs used five different initial starting values for the points to verify consistency over different starting values. This illustrates that even 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 five for the real, experimental HI values would therefore seem significant.

**Simulation Study 2** In this simulation the HI values of the experimentally determined data were randomly permuted within a given HI table. The issue is whether or not such randomly assigned, discrete, HI values can be fit in low dimension. If permuted data could be fit in low dimension, this would vitiate the conclusion that the dimension of immunological shape space, as determined from unpermuted tables of experimental HI values, is low. Clearly, given the the small number of possible discrete HI values, and the size of the data set (14\*19), there will exist a dimension in which it is in fact possible to fit such randomized data. The question is whether this dimension is significantly greater than the five dimensions in which the unpermuted experimental data can be successfully fit.

### **Result 2**

For conciseness, we report results on the Raymond data set [Raymond(1986)] considered earlier. It turns out not possible to fit this data in low dimensions, e.g. dimension five if the relationships between the HI values are broken by randomly

permuting the HI values within a given table. For example, five different sets of initial values for the coordinates of points representing antigens and antisera of the (randomly permuted) data of [Raymond(1986)] were tested in attempts to fit the permuted data in dimension five. All five attempts to fit the permuted data in dimension five were failures. Five such attempts at fitting the same permuted data was performed in dimension ten, and also in dimension fifteen. In dimension ten three of the five initial sets of random values for the coordinates resulted in failure to fit in ten dimensions. In fifteen 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 ten or fifteen dimensions, 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. Discrete HI values relating to distances between  $14 + 19 = 33$  points need deterministic, i.e. non-random relationships, between them for the data to be successfully fit in low dimensions such as five dimensions.

## 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.

The approach presented here 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 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. Alternatively, 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 defined above to sequence space evolution.

Clearly 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.

### **Acknowledgements**

The authors thank the Department of Energy for financial support of this research, and also 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**

- [Perelson (1979)] Perelson A., Oster, G., “Theoretical Studies of Clonal Selection: Minimal Antibody Repertoire Size and Reliability of Self-Non-Self Discrimination”, *J. Theor. Biol.* 81:645-67 (1979)
- [Segel et. al. (1988)] Segel L., Perelson A., “Computations in Shape Space: A New Approach To Immune Network Theory” in “Theoretical Immunology, Part Two”, *SFI Studies in the Sciences of Complexity*, Addison-Wesley (1988)
- [Perelson (1988)] Perelson A., “Towards a Realistic Model of the Immune System” in “Theoretical Immunology, Part Two”, *SFI Studies in the Sciences of Complexity*, Addison-Wesley (1988)
- [DeBoer et. al. (1992)] DeBoer R., Hogeweg P., Perelson A., “Growth and Recruitment in the Immune Network” in “Theoretical and Experimental Insights Into Immunology”, *NATO ASI Series*, Vol. H66, Springer-Verlag (1992).
- [DeBoer et. al. (1992)] DeBoer R., Segel L., Perelson A., “Pattern Formation in One and Two Dimensional Shape Space Models of the Immune System”, *J. Theoretical Biol.* 155:295-333 (1992)

- [Shepherd(1963)] Shepherd R., "Analysis of Proximities as a Technique for the Study of Information Processing in Man", *Human Factors* 5:33-48 (1963)
- [Shepherd(1964)] Shepherd R., "Attention and the Metric Structure of the Stimulus Space" *J. Math. Psychology* 1:54-87 (1964)
- [Kruskal(1964a)] Kruskal J., "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis", *Psychometrika* 29:1-27 (1964)
- [Kruskal(1964b)] Kruskal J., "Nonmetric Multidimensional Scaling", *Psychometrika* 29:115-129 (1964)
- [Borg(1997)] Borg I., Groenen P., "Modern Multidimensional Scaling", Springer Series in Statistics, Springer Press (1997)
- [Braun (1987)] Braun W., "Distance Geometry and Related Methods for Protein Structure Determination from NMR Data" *Quarterly Reviews of Biophysics* 19:115-157 (1987)
- [Edelman(1995)] Edelman S., "Representation of Similarity in 3D Object Discrimination", *Neural Computation* 7:407-422 (1995)
- [CDC] private communication
- [Hillis(1990)] Hillis, D., Moritz, C. (editors) "Molecular Systematics", Sinauer Associates Inc., Sunderland MA (1990)
- [Phylip(1993)] Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle. (1993)
- [Lehmann(1975)] Lehmann E. "Nonparametrics: Statistical Methods Based on Ranks", Holden-Day (San Fransisco), 1975
- [Raymond(1986)] Raymond F, Caton A., Cox N., Kendal A., Brownlee G. "The Antigenicity and Evolution of Influenza H1 Haemagglutinin from 1950-1957 and 1977-1983: Two Pathways from One Gene" *J. Virology* 148:275-287 (1986)
- [Both(1983)] Both G., Sleight M., Cox N., Kendal A. "Antigenic Drift in Influenza Virus H3 Hemagglutinin from 1968 to 1980: Multiple Evolutionary Pathways and Sequential Amino Acid Changes at Key Antigenic Sites" *J. Virology* 48:52-60 (1983)
- [Smith(1999)] Smith D., Forrest S., Ackely D., Perelson A. "Variable Efficacy of Repeated Annual Vaccinations Against Influenza: An in-machina Study" *Proc. Nat. Acad. Sci.* (in press, 1999)