The Perceived Directionality of the Antigenic Evolution of Influenza

Thijs van der Klauw

August 21, 2012

Master Thesis
Mathematical Sciences
Department of Mathematics
Utrecht University

Supervisors:
Martin Bootsma (UU)
Rolf Ypma (RIVM)
Tjibbe Donker (RIVM)

Preface

In this article is described part of my research done to graduate from Utrecht university with a master's degree in mathematics. The decision was made that I would write an article on an interesting topic found in my research. However, I was also required to write down details of the various other topics I looked into during my internship to be able to graduate from the Utrecht University. The appendix below describes these other topics as well as supplemental information to the article.

Abstract

The influenza virus can evade the acquired long term immunity in the human population through changes in its antigenic profile. Understanding the antigenic evolution is crucial to the selection of vaccine strains, which should provide immunity against dominant influenza strain next season. For this purpose, antigenic cartography has been developed to map the changes of the virus over time, using Hemagglutination Inhibition (HI) assay data. Here we present a simple model for the antigenic evolution of the influenza virus, based on a high dimensional random walk in Euclidean space. Using this model we simulate HI assay data which we use to construct antigenic maps through antigenic cartography. Our analysis shows that, although the clustered structure of the data is preserved, the number of dimensions required to describe the data is greatly underestimated. This means that the techniques used in antigenic cartography to assert the dimension required to describe the HI assay datasets does not give conclusive results about the actual dimension of the shape space in which the antigenic evolution takes place. The assumption that the antigenic evolution of influenza is influenced by long term immunity in the human population can therefore not be verified using HI assay data. Our analysis shows that directionless evolution through a high dimensional shape space forms a parsimonious explanation for the observed pattern of influenza evolution.

Introduction

Each year, influenza causes high morbidity and mortality on a global scale, with an estimated 250.000 to 500.000 deaths due to infection [12]. Vaccination against influenza remains the most effective way to prevent a severe infection [1]. However, the virus evolves rapidly on both a genetic and antigenic scale, eluding the vaccine and the natural immunity in the population [2]. To combat this, vaccines are updated about once a year for both the northern and southern hemisphere. The vaccines are based on the currently dominant strains of the virus, as well as strains which are expected to become dominant which differ antigenically substantially from strains currently circulating [3]. A better understanding of the antigenic evolution of influenza could help us to predict future dominant strains better and aid in strain selection for the vaccine.

The main target of the vacinne is the Hemagglutinin surface protein of the influenza virus, which is involved in the binding with host cells. This protein is used frequently as antigenic determinant of the influenza virus [4]. Hemagglutination Inhibition (HI) assays have been developed to assess antigenic similarities between strains of influenza [5] based on this protein. This assay measures the ability of antisera, typically raised in ferrets, to block binding of strains of influenza with red blood cells. For each strain the maximum dilution is determined at which the antiserum is still capable of blocking the binding with such cells.

The results of HI assays can be interpreted using so-called antigenic cartography, a form of multi-dimensional scaling (MDS) [6, 7]. This approach uses the concept of shape space, a theoretical space which described the antigenic characteristics of the influenza virus [8]. Each strain and each antigen is described by a point using an unknown number of coordinates, where each coordinate describes a molecular property that influences the ability of the virus to bind with host cells. Included in these properties are geometric quantities used to describe the shape and size of the binding sites, charge and the ability to form hydrogen bonds. The number of properties required is called the dimension of shape space and is equal to the number of coordinates used to describe each point.

Estimates on the dimension of the shape space range from 2 [7] to 12 [9], based on various approaches. Antigenic cartography, using the similarities between strains and antisera obtained through HI assays, has been used to construct maps of the antigenic evolution in 2, 3, 4, and 5 dimensions. However, it is still uncertain whether these methods are able to correctly estimate the dimension of shape space. Moreover, these estimation may be hindered by the structure of the data inherent to the HI assays.

Here we present a high dimensional model for the antigenic evolution based on minimal assumptions. We use antigenic cartography to reconstruct antigenic maps of our simulated influenza evolution. We assess if the clustering structure of the original dataset is

retained in the reconstructed maps. Furthermore, we estimate the dimension of our simple evolutionary model, to determine whether methods like antigenic cartography are able to correctly identify the dimension of influenza's antigenic evolution.

Materials & Methods

Model

We described antigenic evolution of the influenza virus by a random walk in high dimensional shape space. We described each strain or antiserum using a point of 15 coordinates, resulting in 15 dimensional shape space. We used an iterative process to construct a new strain from the previous one. Then we constructed antisera for a subset of the strains. We then discarded larger distances and introduced error measurements in the generated dataset, as to better resemble the data gathered through HI assays. We then constructed antigenic maps for our datasets in 2, 3 and 4 dimensions. Finally, we analyzed the maps cluster structures and ability to predict unmeasured distances.

To generate the points for the strains of the influenza virus, we used a random walk in high dimensional shape space. This random walk is a discrete time process which produces a new point in each step using the previous point. From point x^t we construct point x^{t+1} by adding a random vector X^t to x^t, where $X^t = (X_1^t, X_2^t, \ldots, X_{15}^t)$ with X_i^t independent identically normally distributed random variables with mean 0 and standard deviation σ^t. We started with $x^0 = \mathbf{0}$, the origin of high dimensional shape space.

Some antigenic changes of the influenza virus are larger than others [7]. We incorporated this in our model by taking either $\sigma^t = 1$ or $\sigma^t = 5$. The choice between 1 and 5 was based on a random vector $B = (B_1, B_2, \ldots, B_{n-1})$ where n is the number of strains we generated. We generated random numbers $J_1, J_2, \ldots, J_l, J_{l+1}$ with $J_k \sim Pois(25)$ for $k = 1, 2, \ldots, l, l+1$ and l such that $\sum_{k=1}^{l} J_k \leq 273$ and $\sum_{k=1}^{l+1} J_k > 273$. We then set $B_i = 1$ for $i = J_1, J_1 + J_2, \ldots, \sum_{k=1}^{l} J_k$ and $B_i = 0$ otherwise. In each step of the iterative process we took $\sigma = 5$ if $B_i = 1$ and else we took $\sigma = 1$. We used this iterative process to construct 273 points.

Next we constructed the points for the antisera. Generally the number of antisera raised to use in the HI assays is far less than the number of available strains [6, 7]. To incorporate this in our model, we selected a subset of strains for which we modeled antisera. To create the subset of strains we generated random numbers $S_1, S_2, \ldots, S_m, S_{m+1}$ with $S_i \sim Pois(3.5)$ for $i = 1, 2, \ldots, m, m+1$, where we picked m such that $\sum_{i=1}^{m} S_i \leq 273$ and $\sum_{i=1}^{m+1} S_i > 273$. We then picked strains $S^j = \sum_{i=1}^{j} S_i$ to be part of the subset for which we generated antisera. Each antiserum was modeled to be the same point as the strain that was used to create it. Because only the inter-point distances between strains and antisera can be measured by the HI assays, we discarded all the strain-to-strain and antisera-to-antisera distances.

Most recent studies on the antigenic evolution of influenza combined data from several

different assays [10, 7]. This results in many missing values between strains and antisera in different tests. Because the HI assays cannot measure large antigenic distances accurately, most of the missing distances in larger HI assay dataset are larger distances. To accomodate for these limitations of combined HI assay dataset, we considered our own upper bound on the distances. We take the upper bound such that 80% of the distances generated by the model are larger than this upper bound. We set each distance larger than the upper bound as a missing value.

As mentioned above, the HI assay cannot measure large distances accurately; the assay returns a threshold value when the distance is too large. This threshold value indicates a minimum inter-point distance for the strain and antiserum in question. To accommodate for threshold values in simulated data we considered each strain separately. For each strain we generate a random number D_i, with $D_i \sim Bin(n, p)$. We took the parameter n to be the number of missing distances for strain i and p to be 3.5 divided by n. For each strain we then randomly selected D_i missing values and replaced them with a threshold value indicating they are larger than the previously set upper bound. We then replaced the D_i largest remaining exact values for this strain and replaced them with missing values. This ensured that 80% of the strain-to-antisera distances remained missing values.

Typical HI assays use only 10 dilution steps in their measurements [5]. To model this, we rescaled the distances in our simulated dataset by dividing it by our upper bound described above. Then, each distance is an element of the interval $[0, 10]$. To accommodate the rescaling done in our threshold values, we set each of them to be >10. The HI assay suffers from interval censoring due to the limited number of dilution steps typically used in the measurements. We simulated this by rounding down 40% of the remaining exact distances.

Antigenic Cartography

Antigenic Cartography uses a multidimensional scaling (MDS) algorithm to construct a configuration of points in $\mathbb{R}^{\hat{d}}$, with $\hat{d}$ a preset target dimension, that best fits a given set of distances between these points. In the case of HI assay data we have a configuration $\mathbf{C}$ that consists of N points for our strains and M points for out antisera. We denote the measured distances, which were generated by our model, between strain i and antiserum j by $D_{i,j}$. The inter-point distances in our configuration between strain i and antiserum j is called $d_{i,j}$. When the measured distance $D_{i,j}$ is an exact value, we wish to ensure that $d_{i,j}$ is close to $D_{i,j}$. If $D_{i,j}$ is a threshold value, we wish to ensure that $d_{i,j}$ is at least as large as the threshold value indicates it should be. If $D_{i,j}$ is missing we don't impose anything on what $d_{i,j}$ should be. We use an error function e to indicate how well a configuration of points fits the given distances. This error function is given below.

$$e(\mathbf{C}) = \sum_{i=1}^{N} \sum_{j=1}^{M} g(D_{i,j} - d_{i,j}) \tag{1}$$

With:

$$g(D_{i,j} - d_{i,j}) = \begin{cases} (D_{i,j} - d_{i,j})^2 & \text{if } D_{i,j} \text{ is exact} \\ (D_{i,j} - d_{i,j} - 1)^2 h(D_{i,j} - d_{i,j} - 1) & \text{if } D_{i,j} \text{ is a threshold} \\ 0 & \text{if } D_{i,j} \text{ is missing} \end{cases}$$

Where $h(x) = 1/(1 + e^{-10x})$. This choice of g and h ensures that, when $D_{i,j}$ is a threshold value, $g(D_{i,j} - d_{i,j})$ only contributes to the error function when $d_{i,j} < D_{i,j} - 1$. We wished to find a configuration which minimizes e. We used the conjugate gradient method for the minimization, where we preformed random restarts to get an approximation of the global minimum of e.

It is difficult to interpret the minimum value of e (1). A lower value indicates a better fit and a higher target dimensions $\hat{d}$ generally decreases the value of e, but it is unclear when we should be satisfied or when we need to further increase $\hat{d}$. Using the map of the optimal configuration found for a target dimension, we measured the map distances between strains and antisera which were missing in the dataset. We compared the map distances with the values we discarded for use by the model. We then took the average of these differences and called this the prediction error. We used this prediction error to judge if an increase in the target dimension gives an improvement in the constructed antigenic map with respect to the dataset used.

Results

Here we present the results of 5 runs, each consisting of 273 strains and associated antisera. In each run we made a distinction between steps which are larger antigenically than other steps. We found that the steps taken with a greater variance for the change per coordinate were larger than those taken with a smaller variance in the change per coordinate (figure 1).

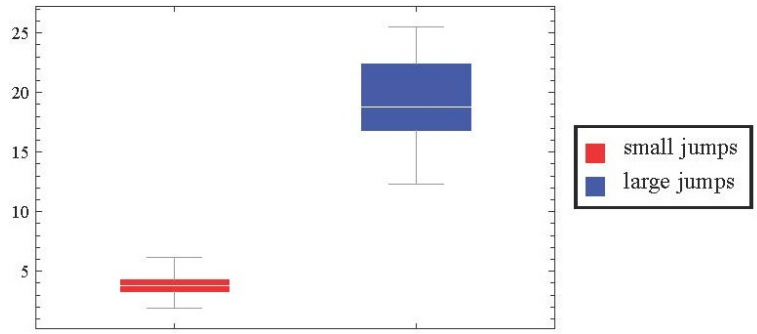

Figure 1: The combined distances between the points subsequently generated of the 5 runs. The small jumps are distances between points where the change per coordinate has $\sigma^t = 1$ and the large jumps are distances between points where the change per coordinate has $\sigma^t = 5$.

The number of clusters and antisera generated varied slightly for each run, with an average of 11.6 and 76.4 respectively. The upper bound on distances we set as measured varied slightly as well (table 1).

run	#clusters	#antisera	average small jump	average large jump	upper bound
1	11	80	3.84	19.40	25.46
2	11	77	3.74	20.02	28.51
3	12	77	3.76	18.34	26.29
4	13	70	3.79	18.37	27.10
5	11	78	3.84	20.33	27.44

Table 1: Various properties of the simulated runs

The antigenic maps presented here are for the first run (figures 2, 3 and 4), but the observations are the same for the antigenic maps of the other runs. There appears to be a 1-dimensional direction in the antigenic maps produced. We also found that the clustering structure of the datasets produced remains intact in the antigenic maps. Furthermore, we found that the MDS algorithm puts antisera and strains apart even if their actual distance

is zero in the antigenic map with target dimension two. The maps produced for target dimensions 3 and 4 are largely the same as the one we produced for target dimension 2, except that we did find a decrease in the distances in these maps between strains and antisera which had actual distance zero (table 2).

run	2 dim map	3 dim map	4 dim map
1	1.86	1.00	0.57
2	1.61	1.00	0.65
3	1.78	0.91	0.40
4	1.81	0.90	0.46
5	1.59	0.89	0.56

Table 2: The average distance between the antiserum and their associated strains in the antigenic maps. The distance between these points is 0 in our model.

We used the prediction error to determine if the resolution of the maps was better in dimensions 3 and 4 than in dimension 2. The prediction error is the average difference between the actual distance of two points on one side and the distance in the antigenic maps produced in target dimension $2, 3$ or 4 on the other. We have shown that the prediction error does not significantly decrease, when we increase the target dimension of our MDS algorithm (table 3).

run	2 dim map	3 dim map	4 dim map
1	13.72	13.87	13.72
2	14.75	13.21	12.82
3	11.09	12.06	12.26
4	13.72	13.58	15.37
5	15.53	14.50	14.05

Table 3: The prediction error per run per target dimension of the antigenic map.

Conclusion & Discussion

We have shown that a simple high dimensional model for the antigenic evolution of influenza can explain the patterns observed in antigenic cartography [7]. We have demonstrated that the resulting antigenic maps of our model show a 1-dimensional direction of the antigenic evolution, even though this direction does not exist in our model. We have also seen that the the clustering structure of the dataset from our model remains intact in the antigenic maps produced. Finally, we have shown that the distance between strains and antisera, with true distance 0, decreases when we increase the dimension of our antigenic map. But the prediction error, the average difference between the true distance of a point and the predicted distance by the map, does not decrease when we increase the target dimension of our antigenic maps.

We used the decrease in prediction errors to determine the required dimension for the antigenic maps produced by antigenic cartography to describe the simulated HI assay datasets. The prediction error did not decrease significantly with increasing target dimensions of the antigenic maps. This leads to the conclusion that 2 dimensions are enough to describe the datasets, while a 15 dimensional shape space was used to generate them. Therefore we conclude that this method of prediction errors cannot be used to determine the true dimension of shape space in which antigenic evolution takes place.

The clustered structure of the modeled HI assay datasets is kept intact when using antigenic cartography, even in two dimensional antigenic maps. The clustering structure did not improve when we increased the dimension of our antigenic maps, therefore we conclude that two dimensions are enough to describe the clustering structure in HI assay datasets using antigenic cartography.

It is generally assumed that the antigenic evolution of influenza on long term immunity [**?**]. In our model each point in the random walk is generated using only the previous point. This means the model used here does not take long term history into account. However, the structure of our simulated antigenic evolution does qualitatively not differ much from the observed influenza evolution. We conclude that it cannot be asserted from HI assay data that long term immunity plays a role in the antigenic evolution, it could just be the high dimension of the shape space in which the evolution takes place.

It has been shown that many external factors, e.g. storage and preparation procedures of the virus strains and variations in animal donors of the red blood cells, influence the test results of HI assays [11]. Furthermore, the larger HI assays used in recent studies combine the data from multiple HI assays with possible differences in accuracy [10, 7]. Therefore, care must be taken when interpreting these results based on HI assay datasets which combine multiple experiments from several laboratories.

HI assays cannot accurately measure large distances between strains and antisera. This means many large values in the HI assay datasets are either missing or reported as a threshold value. We believe the lack of these values is an important factor in the results presented here. A matrix completion algorithm has been constructed to estimate these missing values [12]. We advise more studies should be done to develop and refine techniques to complete the HI assay datasets.

In conclusion, we have shown that a simple high dimensional random walk can explain the results seen in antigenic cartography [7]. This demonstrates that the shape space in which antigenic evolution of influenza takes place could be high dimensional. Furthermore, antigenic cartography can produce antigenic maps which show a 1-dimensional direction of the antigenic evolution, while this direction is not really there. As a result, the data gathered through HI assays cannot prove that antigenic evolution of influenza depends on long term history, while this is generally assumed to be true.

References

[1] WHO, "Influenza Vaccines." http://www.who.int/influenza/vaccines/en/, 2012.

[2] A. Melidou, M. Exindari, G. Gioula, D. Chatzidimitriou, Y. Pierroutsakos, and E. Diza-Mataftsi, "Molecular and phylogenetic analysis and vaccine strain match of human influenza A(H3N2) viruses isolated in Northern Greece between 2004 and 2008," *Virus research*, vol. 145, pp. 220–6, Nov. 2009.

[3] WHO, "Influenza Vaccines (about)." http://www.who.int/influenza/vaccines/about/en/index.html, 2012.

[4] Wikipedia, "Influenza hemagglutinin." http://en.wikipedia.org/wiki/Influenza_(hemagglutinin), 2012.

[5] G. Hirst, "The quantitative determination of influenza virus and antibodies by means of red cell agglutination," *The Journal of experimental medicine*, pp. 49–64, 1942.

[6] A. S. Lapedes and R. Farber, "The geometry of shape space: application to influenza," *Journal of theoretical biology*, vol. 212, pp. 57–69, Sept. 2001.

[7] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. a. M. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *Science (New York, N.Y.)*, vol. 305, pp. 371–6, July 2004.

[8] A. Perelson and G. Oster, "Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination," *Journal of theoretical biology*, vol. 81, no. 4, pp. 645–70, 1979.

[9] X. Du, L. Dong, Y. Lan, Y. Peng, A. Wu, Y. Zhang, W. Huang, D. Wang, M. Wang, Y. Guo, Y. Shu, and T. Jiang, "Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation," *Nature Communications*, vol. 3, p. 709, Feb. 2012.

[10] C. Russell, T. Jones, I. Barr, and N. Cox, "The global circulation of seasonal influenza A (H3N2) viruses," *Science*, vol. 320, no. 2008, 2008.

[11] G. Hirst, "Studies of antigenic differences among strains of influenza A by means of red cell agglutination," *The Journal of Experimental Medicine*, no. 10, 1943.

[12] Z. Cai, T. Zhang, and X.-F. Wan, "A computational framework for influenza antigenic cartography," *PLoS computational biology*, vol. 6, p. e1000949, Jan. 2010.

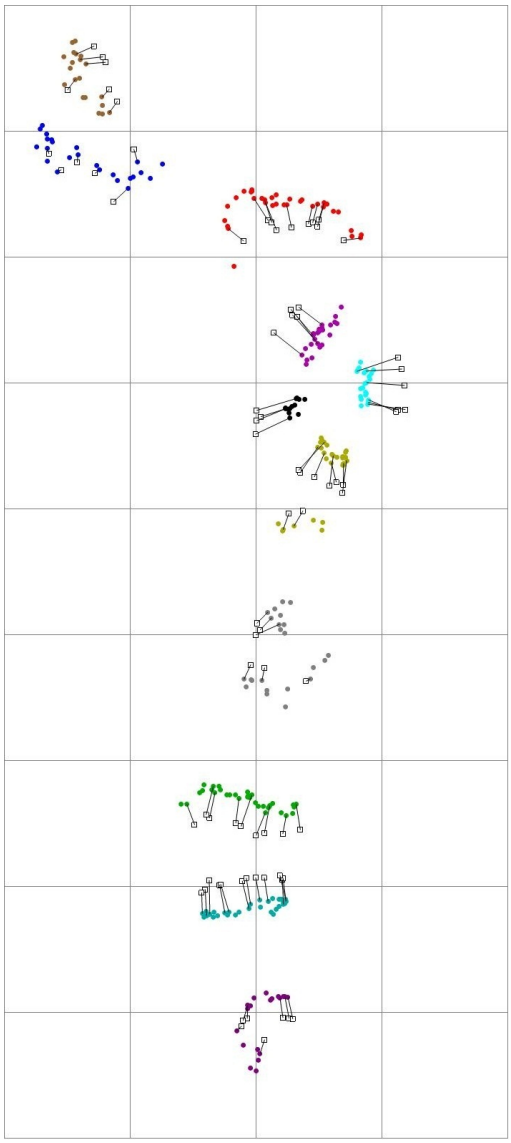

Figure 2: Two dimensional antigenic map for the first run of the model. The dots indicate strains, color coded to show the clustering structure. The square are antisera. A line between a point and square indicate the actual distance between this strain and antiserum is zero.

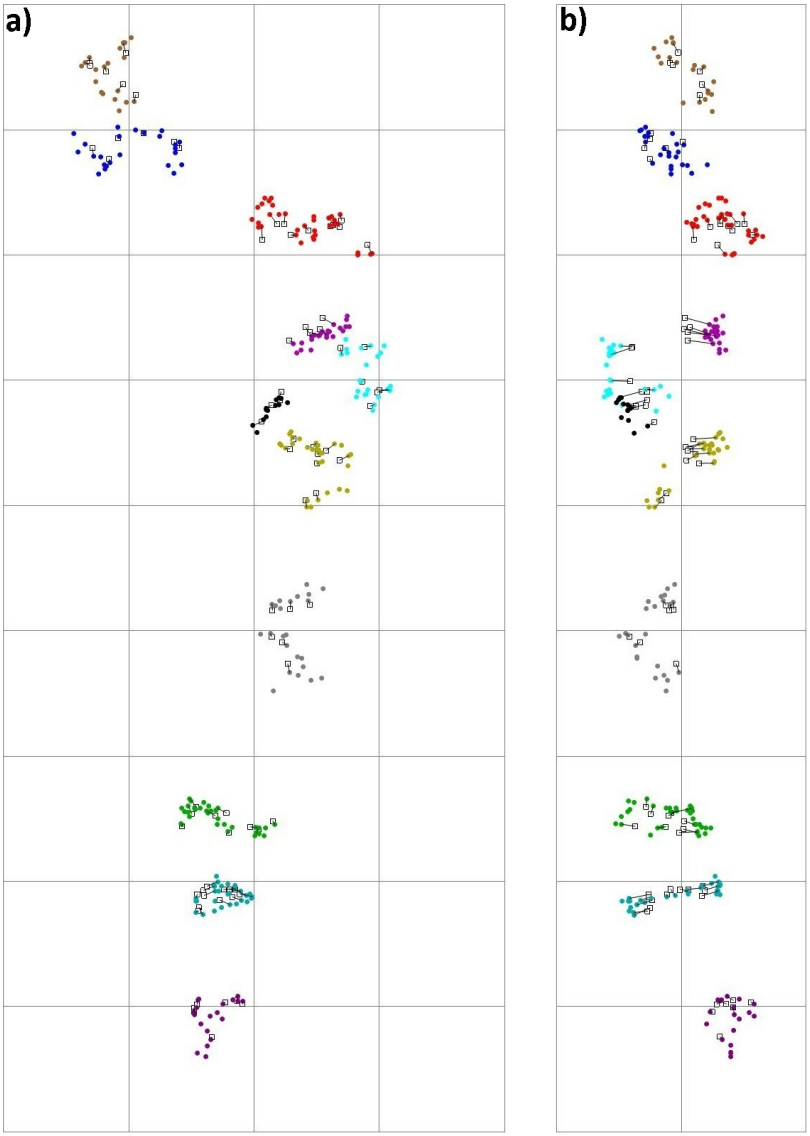

Figure 3: Three dimensional antigenic map for the first run of the model. Principal component analysis was done, where the first two components (a) and the first and the third component (b) are shown.The dots indicate strains, color coded to show the clustering structure. The square are antisera. A line between a point and square indicate the actual distance between this strain and antiserum is zero.

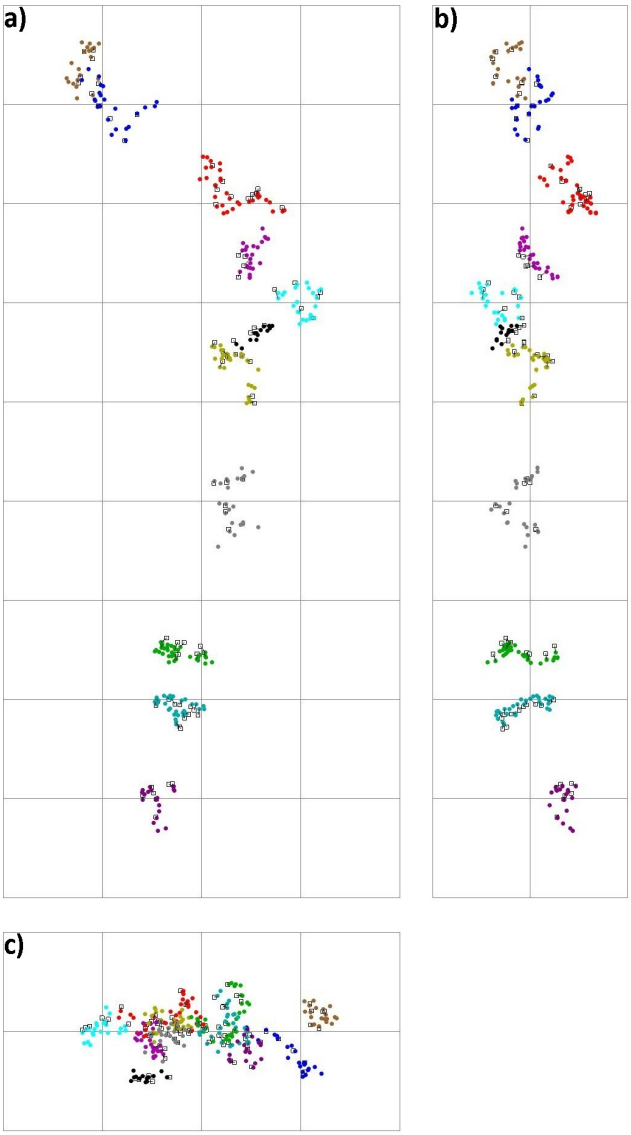

Figure 4: Four dimensional antigenic map for the first run of the model. Principal component analysis was done, where the first two components (a), the first and the third component (b) and the second and the fourth components (c) are shown.The dots indicate strains, color coded to show the clustering structure. The square are antisera. A line between a point and square indicate the actual distance between this strain and antiserum is zero.

Appendix to The Perceived Directionality of the Antigenic Evolution of Influenza

Thijs van der Klauw

August 21, 2012

Contents

Introduction

The influenza virus causes many infections on a yearly basis over the entire world, with an estimated 250.000 to 500.000 deaths due to an infection each year [1]. Even without considering the economic damage caused by people calling in sick or having reduced productivity due to being symptomatically infected, the virus is a major problem for public health. After shedding the influenza virus it is commonly expected that (life) long immunity to the disease is attained, as is the case with many other infections. Unfortunately, reinfection with the virus is very common.

Genetic changes to the influenza virus are usually assumed to be the cause of its ability to avoid natural immunity aquired in the human population. The genetic information of the virus is encoded within RNA molecules, unlike mammels whos genetic material is DNA. RNA is less stable than DNA, causing it to mutate more frequently. Such mutations are known as genetic changes to the virus. These mutations can cause a change in the ability of antibodies raised against the unmutated version of the virus to combat infection by the mutated variant, a change in the antigenic phenotype of the virus.

Over the years, RNA sequencing has been developed in conjunction with phylogenetic tree construction to map the genetic evolution of the influenza virus. These techniques make it possible to see the genetic mutations that occur in the virus over time. However, it is unknown what effect of the near endless posibilities of mutations have on the antigenic phenotype of the virus. Thus we are unable to determine the antigenic changes of the virus over time using these maps.

The virus particles of influenza have so called hemagglutinin proteins on their surface. These proteins play a major role in the binding between the virus particles and host cells, e.g. red blood cells. The antigenic phenotype of a strain of the influenza virus is the ability of various antibodies produced by the immune system to bind with the hemagglutinin proteins found on the virus particle. Bonds formed this way block the ability of the virus particle to bind with the host cells, thus limiting the growth of the virus inside the host.

To evaluate antigenic relations between different strains of the influenza virus (or any virus in general) a binding assay has been developed called hemagglutination inhibition assay (HI assay). Antisera to several strains of the virus are raised in animals, usually ferrets. Then, the ability of the antisera to block binding with red blood cells by the strains is tested for different titers of the antisera. The red blood cells are usually taken from chickens or other avian species.

The results of this assay can be viewed as similarities between different strains of the influenza virus and the antisera specifically raised against these strains. A higher value indicates that the strain and antiserum are antigenically more similar. To interpret these

results we'll use a notion called shape space, a concept suggested as the space in which antigenic evolution takes place [2]. A technique called antigenic cartography, which is based on multi-dimensional scaling, has been developed to produce maps of the antigenic changes of the influenza virus over time. Here we attempt to assess the ability of this technique to accurately determine the dimension of the underlying shape space through constructing antigenic maps.

1 Formulation of the problem

In mathematical terms, shape space is (a subset of) $\mathbb{R}^n$, the standard n-dimensional euclidean space, on which we have the standard euclidean distance. In this space, each strain is described as a point. The number of dimensons n is the number of molecular properties required to describe the ability of the strain to bind with host cells. Antisera raised against strains are descirbed by points in shape space as well. We interpret the similarities found by the HI assay as determinants of the distance between strains and antisera in our shape space.

The abstract, mathematical formulation of the problem now becomes as follows. We have n objects which lie in $\mathbb{R}^d$ with unknown dimension d. We also have a way to determine similarities δ between the objects, where $\delta_{i,j}$ is the similarity between objects i and j. We assume that $\delta_{i,j} = \delta_{j,i}$ and $\delta_{i,i} = 0$. The actual distances between the objects and the measured similarities do not need to be the same, but there is some (usually unknown) relation between them. We assume this relation is strictly monotonic. We want to fit these n points in $\mathbb{R}^{\hat{d}}$. It is always possible to do this for $\hat{d} = n - 1$ regardless of original dimension d. The question at hand is; is it possible to fit the points in a dimension $\hat{d} << n - 1$?

To tackle this problem, Kruskal et al [3] originally formulated the concept of stress. A concept which indicates how well a certain configuration of n points in $\mathbb{R}^{\hat{d}}$ fits the given similarties. Finding the best possible fit in dimension $\hat{d}$ then boils down to minimizing the stress with respect to the coordinates of the n points in $\mathbb{R}^{\hat{d}}$. Depending on the assumptions made about the dissimilarities, there are different ways to tackle the problem; they are described in the following sections.

The underlying goal is to study the antigenic evolution of influenza. We will apply these techniques to data from hemagglutination inhibition (HI) assays, as has been done by Lapedes et al [4] and Smith et al [5]. The goal is to find the dimension of the underlying shape space for antibody-antigen affinity or (an approximation of it in a lower dimension).

2 Metric Multidimensional Scaling (MDS)

Let us assume we know the relation between the results of the HI assays and the actual distances in the shape space between antisera and strains. Kruskal et al [3] described a more general situation followed by the aforementioned concept of stress. This methods can be adopted to this specific case. We have a set of observations from the assay from which we can deduct the target distance $D_{i,j}$ between points i and j. We wish to set a target dimension $\hat{d}$ which is the dimension is which we attempt to find a configuration of points that fits our data. If we take any configuration of points $\mathbf{C} = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^{\hat{d}}$ we wish to measure how well this configuration represents the original n points. We take the squared error between the target distances $D_{i,j}$ (which we know from $\delta_{i,j}$) and the distances of the configuration $d_{i,j}$, where $d_{i,j}$ is the standard Euclidean distance in $\mathbb{R}^{\hat{d}}$ between points x_i and x_j. If we sum over all possible distances we get the following expression:

$$e(\mathbf{C}) = \sum_{j=1}^{N} \sum_{i=1}^{N} (D_{i,j} - d_{i,j})^2 \tag{1}$$

We wish to find the configuration which has the least stress. To do this we minimize this function over all possible configurations of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^{\hat{d}}$. A point in which the minimum is attained is called a best possible fit in this dimension $\hat{d}$. We note that the function e only depends on the inter-point distances in our configuration. Which means that the value of the function will not change if we translate, rotate or reflect our configuration $\mathbf{C}$. We call a configuration a perfect fit if the function e becomes 0. Which only happens if there is a configuration in $\mathbb{R}^{\hat{d}}$ which has exactly the same distances as the target distances. Note that, for the distances we gained from the dissimilarities to make any sense, we need to have that $d_{i,i} = 0$ for all i, the distance from a point to itself is 0. Using this fact we can discard any terms in e_{metric} with $i = j$. Also note that we have assumed $\delta_{i,j} = \delta_{j,i}$. If we combine this with the fact that the euclidean distances obey the same property, we can conclude that $(D_{i,j} - d_{i,j})^2 = (D_{j,i} - d_{j,i})^2$. So our error e (1) becomes $e = \sum_{j=2}^{N} \sum_{i=1}^{j-1} 2(D_{i,j} - d_{i,j})^2 = 2 \sum_{j=2}^{N} \sum_{i=1}^{j-1} (D_{i,j} - d_{i,j})^2$. Where we can discard the factor 2 when we're trying to minimize the function. The function we wish to minimize becomes:

$$e_{\text{metric}}(\mathbf{C}) = \sum_{j=2}^{j=N} \sum_{i=1}^{j-1} (D_{i,j} - d_{i,j})^2 \tag{2}$$

We note the following. First, it is a summation of nonnegative terms, so the entire function is nonnegative. If we have any set of points in $\mathbb{R}^d$ for some fixed d and we use their inter-point distances to construct e_{metric} for target dimension d from it, then this exact set of points will obviously result in $e_{\text{metric}} = 0$. But this does not need to be the case when we construct e_{metric} for target dimension $\hat{d} < d$. It is difficult to interpret this value with respect to the minimum value of the error function found in other target

dimensions. One might wonder why it would be important to interpret error function values greater than 0, since we wish to obtain the actual shape space in which antigenic evolution of influenza takes place, so we wish to have a value of 0. But we can not hope to find a correct answer even if we use the dimension for which the error is 0, since the HI assays are crude and carry with them many measurement errors [6]. It might even be possible that the measurements introduce distances which do not satisfy the triangle inequality. This means a perfect fit can never be found, as no set of points can satisfy these target distances.

3 Ordinal MDS

The method of metric MDS, as described in section 2, fails when we can no longer assume we know the exact relation between the similarities and the actual distances. We still assume that this relation is monotonic. Which is a much more plausible and general assumption, more suitable for our data. When using the method of ordinal MDS, we only make this assumption. Note that a larger similarity between two points indicates that the distance between the two points should be smaller, so we get that $\delta_{i,j} > \delta_{k,l}$ implies $d_{i,j} < d_{k,l}$. To get this we apply the following trick to our data. To each combination i, j of the n points, we assign a number α, which is the order, when ordering from highest to lowest, of $\delta_{i,j}$ among all the observed similarities. For example, if $\delta_{i,j}$ happens to be the largest among all observed similarities, we take $\alpha = 1$, if it is the second largest then $\alpha = 2$. While if $\delta_{i,j}$ is the smallest then $\alpha = N$ where N is the total number of similarities measured. By construction we now have that $\delta_{\alpha+1} < \delta_\alpha$, which implies that $d_{\alpha+1} - d_\alpha > 0$ for all $\alpha \in \{1, 2, \ldots, N-1\}$. Lapedes [4] suggests to use the following function in his article.

$$e_{\text{ordinal}}(\mathbf{C}) = -\sum_{\alpha=1}^{N-1} \log\left(g(d_{\alpha+1} - d_\alpha)\right) \tag{3}$$

With g a continuous approximation of the indicator function on $\mathbb{R}_{>0}$, also called a squashing function. We want a continuous approximation as we will use a minimization technique which requires a differentiable function, I use $g(x) = 0.5(1 + \tanh(x))$. This squashing function rapidly approaches 1 for positive values and 0 for negative values. Since $g(x) \in (0, 1)$ for all x we have that $-\log(g(x)) \in (0, \infty)$ for all x. If $d_{\alpha+1} - d_\alpha \gg 0$ we get that $-\log(g(d_{\alpha+1} - d_\alpha))$ approaches zero. So the entire function approaches zero if $d_{\alpha+1} \gg d_\alpha$ for all $\alpha \in \{1, 2, \ldots, N\}$. It is quite evident that this function is much more complex, as its form depends on the ordering of the measured similarities between the points.

Note that we need not have a very small value of e if the points in our configuration do obey $d_{\alpha+1} > d_\alpha$ for all $\alpha \in \{1, 2, \ldots, N\}$. Since the value of $-\log(g(d_{\alpha+1} - d_\alpha))$ need not be very small if $d_{\alpha+1}$ and d_α only differ by small numbers. So we need to ensure that we have a different method for checking if a configuration does satisfy the conditions $d_{\alpha+1} > d_\alpha$ for all $\alpha \in \{1, 2, \ldots, N\}$, rather than trying to get $e_{ordinal}$ to become 0.

Note that this function can only be constructed if there are no ties in the ordering of the δ's. If there are a decision must be made which of the two will be assigned the smaller α. If we take a set of randomly generated points which have a continuous distribution for their interpoint distances, the chance of getting two distances which are exactly the same is 0. But the data generated by HI assays does not have this nice property. Typically a two-fold dilution is made in each step of the assay, where the result for a given strain/antiserum combination is the highest dilution at which the antiserum was still able to block binding of the virus with red blood cells. It should be clear that this

results in many similarities being the same.

3.1 Binned data

What we have in reality from our HI assays is not a value of the true similarity of a strain and antiserum, but an indicator between which possible values it lies. The similarity is the maximum dilution at which the antiserum still blocks binding of the strain with RBC's. So the result for each strain and antiserum combination indicates a range of possible values for the true similarity, which we will call a bin. First, we assume there is a bin for each possible value of the similarities, so we assume the assay can at least assign a value to each possible similarity. We also assume that the range of values any bin indicates is an interval which does not overlap with the range of values indicated by any other bin. We do this to ensure that we can still say something about the order of similarities in different bins. In the case of HI assay data the intervals are closed on the left and open on the right. The result is that if we look at the range of possible similarities, which is $\mathbb{R}_{>0}$, that we have a (usually finite) set B of values in $\mathbb{R}_{>0}$ which are the bin values. Because of our assumptions on these bin values, we can order the bin values according to which range they indicate. Since there are only a finite number of similarities measured, the number of different bins is finite and bounded by the number of different similarities we measure. An example of an arbitrary binning of the positive real line can be seen in figure 1. The value reported for each bin by the HI assay is the minimum of the interval, which exists because the intervals are closed on the left.

Figure 1: An example of a binning of the positive reals. With the the red lines indicating the edges of the bins. A similarity of value 1.9 will return the value of bin 2 when measured.

Because we knew the total order of the similarities in our previous versions of the problem, we had that $\delta_\alpha > \delta_\beta$ and $\delta_\beta > \delta_\gamma$ implies $\delta_\alpha > \delta_\gamma$. In the case of bins we can have the following situation. Assume δ_α and δ_β lie in the same bin and that δ_γ lies in a lower bin. If we only require that $\delta_\gamma < \delta_\alpha$ we know nothing of the relation between δ_γ and δ_β. So we need to require that $\delta_\gamma < \delta_\beta$ as well. As mentioned before, the original formulation of our method of ordinal MDS fails in the case of ties, which is exactly what happens when we have bins. Rather than completely ordering the distances using the ordering of the similarities as done before, we now use the ordering, from large to small, among the bins of the similarities to bin the distances. Now we want to ensure that all the distances in our bins are smaller than any distances in the subsequent bin. Previously we had one inequality we wanted to satisfy directly per distance, while the other inequalities would then be satisfied automatically. When we have bins and a given

distance, we need an inequality for each distance in the next bin. It is easy to see that the amount of inequalities we need to satisfy quickly increases when the number of bins is low compared to the number of similarities. Unfortunately, this is typically the case in the assays used.

While Lapedes et al use binned data for the ordinal MDS algorithm, as suggested in [4], no clear explanation is given for the exact modification he made to the algorithm to accommodate for the existence of these bins in the data. I suggest the following modification to $e_{ordinal}$:

$$e_{\text{ord,bins}}(\mathbf{C}) = -\sum_{\alpha=1}^{N} \sum_{\beta \in B_\alpha} \log\left(g(d_\beta - d_\alpha)\right) \tag{4}$$

Here we ordered the distance bins using the ordering of the similarity bins, as we assumed is possible. We take B_α the set of distances in the next bin (which depends on α). We define $B_\alpha = \emptyset$ when α lies in the last bin. If we indeed have that $d_\beta \gg d_\alpha$ for all β in the next bin and we have this for all $\alpha \in \{1, 2, \ldots, N\}$ we get that $e_{ord,bins}$ approaches zero.

There is another possible workaround which runs into other technical difficulties. If we ensure that each distance is smaller than the smallest distance in the next bin, it is obviously smaller than all the distances in this bin. This reduces the amount of inequalities we want to satisfy back to $N-1$. To reduce the amount of inequalities even further we could only do this for the largest distance in each bin and observe that all the previous inequalities still hold. The amount of remaining inequalities would be the number of bins minus 1. The error function to minimize now becomes:

$$e_{ord,ext}(\mathbf{C}) = -\sum_{k=1}^{\#bins-1} \log\left(g(\min_{\beta \in B_{k+1}} d_\beta - \max_{\alpha \in B_k} d_\alpha)\right) \tag{5}$$

The problem with this function is that it becomes rather hard to minimize because it includes a minimum and a maximum. The function depends on the ordering of the distances of different configurations, whereas (4) did not. Since the suggested method of conjugate gradient minimization (see section 5) uses differentiation on the function this becomes a problem: the form of the gradient of $e_{ord,ext}$ depends on the order of the distances in your current configuration. Whereas, in the previous methods, a general form of the gradient could be computed depending on the coordinates of the configuration alone. I don't believe the advantage gained by (vastly) reducing the amount of inequalities we need to satisfy outweighs the technical difficulties posed by calculating this gradient.

A third possible solution was suggested to me by Rolf Ypma. It ultimately failed to work, but I wish to mention it anyway as I found its complications rather interesting. Rolf's suggestion was to order the distances in each bin randomly before each computation

and then use this order to return to the case of ordinal MDS without bins. Iterating this process over many random orderings and taking the configuration for which the minimum value is the lowest, could in theory result in the actual desired best fit in a dimension. At least the result would give an upper bound for the lowest dimension in which a fit without errors is still possible. However, it is possible to generate points in dimension d, bin them and then rearrange them in each bin such that the dimension $\hat{d}$ for which no errors in ordering occur is actually larger than d. Consider the following example:

- We generate 4 points on the real line, so the original dimension is 1.
- The points are $x_1 = 0, x_2 = 1, x_3 = 1.1, x_4 = 10$
- The corresponding distances are $d_{1,2} = 1, d_{1,3} = 1.1, d_{1,4} = 10, d_{2,3} = 0.1, d_{2,4} = 9, d_{3,4} = 8.9$
- Take 4 bins constructed such that:
  - Bin 1 contains $d_{2,3} = 0.1$
  - Bin 2 contains both $d_{1,2} = 1$ and $d_{1,3} = 1.1$
  - Bin 3 contains both $d_{2,4} = 9$ and $d_{3,4} = 8.9$
  - Bin 4 contains $d_{1,4} = 10$
- We order the distances in bin 2 by setting $d_{1,2} < d_{1,3}$ and the distances in bin 3 by setting $d_{2,4} < d_{3,4}$

If we want to fit these 4 points into 1 dimension we immediately see that $d_{1,4}$ has to be the largest distance, since it is the only distance in the last bin. So points x_2 and x_3 will lie in between x_1 and x_4. From the order we set in bin 2 we can conclude that x_2 should lie closer to x_1 than x_3. From the order we set in bin 3 we conclude that x_2 should lie closer to x_4 than x_3. This is clearly not possible. We conclude that we cannot fit these 4 points into 1 dimension without violating the order we set. Hence we conclude that the method could give a large error even in the original dimension in which the points lie.

3.2 The number of errors for any given configuration

As mentioned in the previous section, we need a different way to measure the goodness of fit of a configuration of points to the set of given similarities, other than the function value of $e_{ord,bins}$. Since the relation between the similarities and distances is not known, we cannot use the original distances to give some form of error as we did with the metric MDS algorithm. We did assume that the relation between distances and similarities is monotonic. We used this relation to determine the ordering between the distance bins. This ordering is used to calculate the amount of errors of a given configuration with respect to this ordering. For each distance d we check if the distances in all the subsequent

bins are larger. If we find a distance in a subsequent bin which is smaller we count an error. The total number of errors found is now our measure for goodness of fit of the configuration to the similarities.

Let us show how this method works through an example. Assume we have 4 points and thus 6 distinct similarities. Assume they are ordered in the following way; $\delta_{1,2} > \delta_{1,3} > \delta_{1,4} > \delta_{2,3} > \delta_{2,4} > \delta_{3,4}$. Finally, assume we have two configurations where the distances obey $d_{3,4} < d_{1,2} < d_{1,3} < d_{1,4} < d_{2,3} < d_{2,4}$ and $d_{1,2} < d_{1,3} < d_{1,4} < d_{2,3} < d_{3,4} < d_{2,4}$ respectively. In the configuration our method counts 5 errors, since $d_{3,4}$ is smaller than each of the other distances, while $\delta_{3,4}$ lies in the last bin. In the second configuration our method counts 1 error, since $d_{3,4} < d_{2,4}$ while $\delta_{2,3} > \delta_{2,4}$.

4 Peculiarities of HI assay data

So far we have described conventional MDS algorithms, adopted for binned values in the ordinal case, to tackle to the problem of fitting n objects in $\mathbb{R}^{\hat{d}}$ with target dimension $\hat{d}$. When trying to adapt the algorithms to construct antigenic maps from HI assay data, we run into other problems specific for the HI assay data. Here we describe these problems, divided into two categories.

4.1 Missing values

Typically a missing value is not problematic. We just ignore this similarity for any of the algorithms and ignore it's associated distance in any configuration of points. We do this because we cannot say anything about the missing value. We also face the fact that many large HI assay tables are constructed by combining results from many smaller assays done over the years. Generally the data has many similarities measured within what we call clusters, groups of strains/antisera which lie close to each other genetically (and presumably antigenically as well). We have to consider that the assay suffers from a sensitivity bound, which means it is hard to measure the similarities between clusters anyway.

4.2 Threshold values

The assay has a sensitivity bound. Beyond this bound, the assay can only report that the similarity is smaller than a given value, the result is typically of the form <10. The ordinal algorithm can work with this, provided the threshold similarities are in the smallest or largest bin. This is not always the case, since most larger data sets of HI assays combine data gathered over many years, where the sensitivity bounds might change over time. If the tested similarity between two points is <40 we do know that the distance between these points should be larger than the distance between points with measured similarity $100, 50$ or 40. But we cannot compare it to a distance between points with measured similarity of 10. This complicates the ordinal MDS algorithm. If we consider (4) we see that we need to adapt the concept of similarities in the next bin. We order the bins once more on their value but only consider bins that have an exact similarity rather than a threshold similarity. We only know the upper bound on the threshold similarity. Therefore we cannot compare it with any bins with an exact or threshold similarity which is lower. But we based our function e_{ordinal} on the similarities which we can pinpoint as being the next in the ordering. So we should omit our distances associated with threshold similarities in the summation. We do know which similarities are larger than our threshold values however. For any bin with a threshold similarity we associate it to the bin with the lowest exact similarity which is still greater than our threshold similarity. We get that (4) becomes (figure 2):

$$e_{\text{ord,thres}}(\mathbf{C}) = -\sum_{\alpha \in A} \sum_{\beta \in B_\alpha} \log\left(g(d_\beta - d_\alpha)\right) \tag{6}$$

Where A is the set of distances in bins with exact values. B_α is defined as the set of distances in the next bin in the ordering and any distances in a bin with threshold value associated with the bin α lies in. Note that B_α need not be empty for α in the last exact bin. There can be bins with a smaller threshold value. $e_{\text{ord,thres}}$ will approach zero if the ordering we had before between the exact bins is not violated and if all distances in thershold bins are larger than distances in exact bins with a larger value than the threshold similarity.

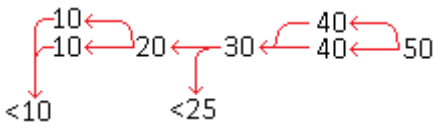

Figure 2: Arbitrarely picked binned similarity values with (some) threshold bins. The arrows indicate the respective sets B_α in (6) for each α. There are no arrows going out of the threshold values, as they do not have a set B_α associated with them.

The metric MDS algorithm has more severe problems with threshold values, since the form of e_{metric} given by (2) cannot deal with values which do not have a specific target distance. When we deal with a threshold similarity, we change the term $(D_{i,j} - d_{i,j})^2$ in the summation of e_{metric} to $(D_{i,j} - d_{i,j} - 1)^2 h(D_{i,j} - d_{i,j} - 1)$. With $h(x) = 1/(1 + e^{-10x})$, another (continuous) approximation of the indicator function on $\mathbb{R}_{>0}$ and $D_{i,j}$ the transformation of our threshold value to a distance. This choice ensures that there is only a (large) contribution to the total error when $D_{i,j} - d_{i,j} - 1 > 0$. So when $d_{i,j} < D_{i,j} - 1$, i.e. when the distance between points i and j is lower than the minimum distance indicated by the threshold value, with a correction of -1 due to binning. Our error function now becomes

$$e_{\text{met,thres}}(\mathbf{C}) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \rho(D_{i,j} - d_{i,j}) \tag{7}$$

With $\rho(D_{i,j} - d_{i,j}) = (D_{i,j} - d_{i,j})^2$ if $\delta_{i,j}$ lies in an exact bin and $\rho(D_{i,j} - d_{i,j}) = (D_{i,j} - d_{i,j} - 1)^2 h(D_{i,j} - d_{i,j} - 1)$ if $\delta_{i,j}$ lies in a threshold bin and $h(x) = 1/(1 + e^{-10x})$.

5 Algorithms and their complexity

In previous sections methods for metric and ordinal MDS were discussed. With a distinction between (fully) accurate data and data with binning and thresholds. I have posed before that the complexity of these algorithms is high (and thus their computational time is as well). Let us continue with giving an explanation for this claim.

As seen in previous sections, each method boils down to choosing a dimension $\hat{d}$ in which we want to try and find a fit. We get an error function depending on the coordinates of any configuration of n points in $\mathbb{R}^{\hat{d}}$. We continue with minimizing this error function to obtain a configuration for which the minimum is attained. We call this configuration the best fit for this dimension $\hat{d}$.

5.1 Conjugate Gradient Minimization

The minimization of our error function e required to find the best configuration is typically done through the conjugate gradient method. For this algorithm we need the gradient of the function we wish to minimize. We start out in an initial configuration $\mathbf{C}_0$ and calculate $x_0 = -\nabla e(\mathbf{C}_0) = s_0$, the steepest descent direction of e in $\mathbf{C}_0$. We then find the α_0 which minimize $e(\mathbf{C}_0 + \alpha x_0)$, where we use an inexact line search. Then we set $\mathbf{C}_1 = \mathbf{C}_0 + \alpha_0 x_0$. Then we loop the following steps until we either find a gradient which has norm less then ϵ or n iterations have passed, where ϵ and n are preset parameters.

- Calculate the steepest descent direction $x_n = -\nabla e(\mathbf{C}_n$.
- Compute $\beta_n = \max\left[0, \frac{x_n^T(x_n - x_{n-1})}{x_{n-1}^T x_{n-1}}\right]$.
- Set $s_n = x_n + \beta_n s_{n-1}$.
- Find α_n which minimizes $e(\mathbf{C}_n + \alpha_n s_n)$.
- Set $\mathbf{C}_{n+1} = \mathbf{C}_n + \alpha_n + s_n$.

Where the α_n in each step is calculated through an inexact line search. We apply this method to various cases we described in sections 2, 3 and 4. Now we wish to say something about the complexity of each of the different cases.

5.2 Metric MDS

For the metric MDS algorithm we assume to have the exact relation between the similarities and the distances in the underlying shape space. Our function e_{metric} is given by (2). Let us start with calculating the gradient of this function as we need this gradient for our minimization process. Taking the derivative to coordinate l of point k we get the following:

$$\begin{aligned}\frac{\partial e_{metric}}{\partial x_{k,l}} &= \frac{\partial}{\partial x_{k,l}} \sum_{i<j} (D_{i,j} - d_{i,j})^2 \\ &= \sum_{i<j} \frac{\partial}{\partial x_{k,l}} (D_{i,j} - d_{i,j})^2 \\ &= \sum_{i<j} -2(D_{i,j} - d_{i,j}) \frac{\partial d_{i,j}}{\partial x_{k,l}}\end{aligned}$$

Where $\frac{\partial d_{i,j}}{\partial x_{k,l}}$ is 0 when $k \neq i, j$ and $\frac{(x_{k,l} - x_{j,l})}{d_{i,j}}$ or $\frac{(x_{k,l} - x_{i,l})}{d_{i,j}}$ when $k = i$ or $k = j$ respectively. We can rewrite the above to:

$$\frac{\partial e_{metric}}{\partial x_{k,l}} = \sum_{j=1, j\neq k}^{j=n} -2(D_{k,j} - d_{k,j}) \frac{x_{k,l} - x_{j,l}}{d_{k,j}} \tag{8}$$

Our sum runs over $n-1$ different values and our gradient has $\hat{d}n$ components (recall that $\hat{d}$ is the target dimension). So we need to calculate $n^2\hat{d}$ terms in the summation. The summations in e_{metric} are both bounded by n, so for this function we need to calculate at most n^2 terms in the summations.

5.3 Metric MDS with binning, thresholds and missing values

The metric MDS algorithm can easily be adapted to deal with bins. We simply pick each target distance $D_{i,j}$ to be some value in the bin, usually the middle of the interval the bin covers. But, as discussed in section 5, the MDS algorithm does run into problems when incorporating threshold values. The adapted function $e_{met,thres}$, see (7), is severaly more complex than our previous function. When calculating the gradient of this function $\rho(D_{i,j} - d_{i,j}) = (D_{i,j} - d_{i,j})^2$ when the similarity between i and j is not a threshold value, which is exactly the same as before. When the similarity associated with $D_{i,j}$ is missing, we get that $\rho(D_{i,j} - d_{i,j}) = 0$ and the derivative becomes zero. If it is a threshold value however, we have that $\rho(D_{i,j} - d_{i,j}) = (D_{i,j} - d_{i,j} - 1)^2 h(D_{i,j} - d_{i,j} - 1)$ with $h(x) = 1/(1 + e^{-10x})$. Once more we take the derivative with respect to coordinate l of point k. We note that $\frac{\partial \rho(D_{i,j} - d_{i,j})}{x_{k,l}} = 0$ if $k \neq i, j$. Let us calculate $\frac{\partial \rho(D_{i,j} - d_{i,j})}{x_{i,l}}$, for some arbitrary l, when the similarity between i and j is a threshold value.

$$\begin{aligned}\frac{\partial\rho(D_{i,j}-d_{i,j})}{x_{i,l}} =& \frac{\partial}{x_{i,l}}((D_{i,j}-d_{i,j}-1)^2\cdot h(D_{i,j}-d_{i,j}-1))\\ =& \frac{\partial(D_{i,j}-d_{i,j}-1)^2}{x_{i,l}}\cdot h(D_{i,j}-d_{i,j}-1)\\ &+\frac{\partial h(D_{i,j}-d_{i,j}-1)}{x_{i,l}}\cdot(D_{i,j}-d_{i,j}-1)^2\\ =& -2(D_{i,j}-d_{i,j}-1)\cdot\frac{x_{i,l}-x_{j,l}}{d_{i,j}}\cdot h(D_{i,j}-d_{i,j}-1)\\ &+h'(D_{i,j}-d_{i,j}-1)\cdot\frac{x_{i,l}-x_{j,l}}{d_{i,j}}\cdot(D_{i,j}-d_{i,j}-1)^2\end{aligned}$$

Where $h'(x)=\frac{dh}{dx}(x)$. We have that $h'(x)=\frac{d}{dx}(1/(1+e^{-10x}))=\frac{10e^{-10x}}{(1+e^{-10x})^2}$. Since $D_{i,j}=D_{j,i}$ and $d_{i,j}=d_{j,i}$ we omit the case $k=j$, as it is analogue to the case we just did. We get the following term for our partial derivative:

$$\frac{\partial e_{met,thres}}{\partial x_{k,l}}=\sum_{j=1,j\neq k}^{j=n}\rho'(D_{k,j}-d_{k,j}) \tag{9}$$

Where we take $\rho'(D_{k,l}-d_{k,l})$ to be $\frac{\partial\rho}{x_{k,l}}(D_{k,l}-d_{k,l})$, which we calculated above in all of the cases. The number of terms in the summation remains the same, but we note that each individual term becomes more complex.

5.4 Ordinal MDS

In the case of ordinal MDS we only know the relative order of the distances between our n points. In this case we assume that there are no ties in this order, as this makes matters more complicated, more on that below. Our error function $e_{ordinal}$ is given by (3). We again need a gradient so let us calculate the derivative to an arbitrary coordinate $x_{i,j}$. Note that we use the ordering given by the α's as they were introduced in section 4. We get the following:

$$\begin{aligned}\frac{\partial e_{ordinal}}{\partial x_{i,j}} &= \frac{\partial}{\partial x_{i,j}}\left(-\sum_{\alpha=1}^{N-1}\log\left(g(d_{\alpha+1}-d_\alpha)\right)\right)\\ &= -\sum_{\alpha=1}^{N-1}\frac{\partial}{\partial x_{i,j}}\log\left(g(d_{\alpha+1}-d_\alpha)\right)\\ &= -\sum_{\alpha=1}^{N-1}\frac{1}{g(d_{\alpha+1}-d_{alpha})}\cdot\frac{\partial g(d_{\alpha+1}-d_\alpha)}{\partial x_{i,j}}\\ &= -\sum_{\alpha=1}^{N-1}\frac{g'(d_{\alpha+1}-d_\alpha)}{g(d_{\alpha+1}-d_{alpha})}\cdot\frac{\partial(d_{\alpha+1}-d_\alpha)}{\partial x_{i,j}}\end{aligned}$$

Where $g'(d_{\alpha+1} - d_\alpha)$ is the derivative of g in the point $d_{\alpha+1} - d_\alpha$. Recall that $g(x) = 0.5(1 + \tanh x)$, so $g' = 0.5(1 - \tanh^2 x)$. Then $\frac{g'}{g}$ becomes $\frac{1-\tanh^2 x}{1+\tanh x} = 1 - \tanh x$. Which we can denote by $h(x)$. All that remains is to determine $\frac{\partial d_\alpha}{\partial x_{i,j}}$ and $\frac{\partial d_{\alpha+1}}{\partial x_{i,j}}$. They are 0 if distance α (respectively $\alpha + 1$) does not depend on point i. If distance α (respectively $\alpha + 1$) does depend on point i then the derivative becomes $\frac{x_{i,j} - x_{k,j}}{d_\alpha}$ (respectively $\frac{x_{i,j} - x_{k,j}}{d_{\alpha+1}}$), where $k \neq i$ is the other point distance d_α (respectively $d_{\alpha+1}$) depends on. This summation can be rather nicely rewritten provided that distances 1 and $NM - 1$ do not depend on point i. We get:

$$\sum_{\alpha \in I_i} \frac{(x_{i,j} - x_{k_\alpha,j})}{d_\alpha}(h(d_{\alpha+1} - d_\alpha) - h(d_\alpha - d_{\alpha-1}))$$

Where I_i is the set of all distances depending on point i and k_α is the other point distance d_α depends on. If either d_1 or d_N depends on point i we have a problem, since then either $d_{\alpha-1}$ or $d_{\alpha+1}$ does not exist. We can remedy this by adding distances d_0 and d_{N+1} to our set of distances and setting $d_0 = -\infty$ and $d_{N+1} = \infty$. We then also extend our function $h(x)$ by $h(\infty) = 0 = \lim_{x\to\infty} 1 - \tanh x$ and $h(-\infty) = 2 = \lim_{x\to-\infty} 1 - \tanh x$. Using these properties, if we have that d_1 or d_N depends on point i, we see that $d_{N+1} - d_N = 0$ (respectively $d_1 - d_0 = 0$) and we'll get that the term $h(d_{N+1} - d_N)$ (respectively the term $h(d_1 - d_0)$) vanishes. We conclude that:

$$\frac{\partial e_{ordinal}}{\partial x_{i,j}} = \sum_{\alpha \in I_i} \frac{(x_{i,j} - x_{k_\alpha,j})}{d_\alpha}(h(d_{\alpha+1} - d_\alpha) - h(d_\alpha - d_{\alpha-1})) \tag{10}$$

The sum in each component of the gradient runs over at most $n - 1$ different values. So we need at most $n - 1$ terms for each part of our gradient. Since there are $n\hat{d}$ different coordinates and thus components of the gradient, the total gradient can then consists of at most $\hat{d}n(n - 1)$ terms.

5.5 Ordinal MDS with binning, thresholds and missing values

In the case of binning and thresholds, we assume that we don't know the exact ordering of the distances between the points. But we have some information on their position. The error function $e_{ord,thres}$ in this case is given by (6). We note that this function is essentially the same as the function we used without binning, only our summation index changed. We don't consider the next distance according to the ordering of our similarities, but we consider all the distances in the next bin according to the ordering of the bins. The derivative of the term inside the summations is the same as above, so we conclude that our gradient is the following:

$$\frac{\partial e}{\partial x_{i,j}} = \sum_{\alpha \in I_i} \sum_{\beta \in B_\alpha} \frac{(x_{i,j} - x_{k_\alpha,j})}{d_\alpha}(h(d_\beta - d_\alpha)) - \sum_{\beta \in B'_\alpha} \frac{(x_{i,j} - x_{k_\alpha,j})}{d_\alpha}(h(d_\alpha - d_{\beta'})) \tag{11}$$

In the above equation we have I_i the set of distances which depend on point i. B_α and B'_α the sets of distances which are in the next (respectively the previous) bin combined with any values in threshold bins associated with our distance d_α as described in section 4. k_α is the other point distance d_α depends on.

Since we do not know the number of distances in each bin an upper bound for the number of terms in the gradient becomes larger. We do know that there are a total of N similarities, which is bounded by $n(n-1)/2$. We know that each distance is considered twice, once for each point it depends on. We also know that whenever a distance is considered, all the distances in the previous and next bins are considered. We wish to know what the worst possible distribution of distances over the bins is for the number of terms in our gradient.

Proposition 1. *The largest amount of terms in the gradient occurs when the N similarities are spread over 2 bins with $a = \lfloor N/2 \rfloor$ in the first bin and $b = \lceil N/2 \rceil$ in the second bin. The number of terms is then given by $4\hat{d}ab$*

Proof: We know that each distance is used $2\hat{d}$ times, since it depends on 2 points each with $\hat{d}$ coordinates. So the total number of terms in the gradient is the sum of the number of distances in the previous and next bins for each distance multiplied by $2\hat{d}$. Thus, to find the worst case scenario, we need to maximize this number. Since threshold values are only associated with at most one bin and exact values have constraints by distances in at most two other bins, we can assume without loss of generality that all our similarities are exact values.

Let us consider any distribution over the bins which has distances in more than 2 bins. We can then shift the distance in the last bin to 2 bins lower. Since these distances are now in the previous bin with respect to distances in the second last bin while they were in the next bin before, the number of terms does not decrease. We conclude that the new distribution over the bins guarantees at least as many terms in the gradient. We can iteratively repeat this process to conclude that the worst case scenario has distances in at most 2 bins.

When we are left with distances in two bins, we have the following. We take each distance $2\hat{d}$ times and count the number of distances in the other bin. We see that the total number of terms is given by $2\hat{d}a \cdot b + 2\hat{d}b \cdot a = 4\hat{d}ab$ with a and b the number of distances in the first and second bin respectively. We wish to maximize this number,

so we can discard the constant factor $4\hat{d}$ for now. So we wish to maximize ab with the constraint that $a+b=N$ and both a and b are positive integers. First note that because of symmetry we can assume that $a \leq b$. We now claim that $a = \lfloor N/2 \rfloor$ and $b = \lceil N/2 \rceil$ is the maximal solution. To see this assume we have $x = a - c$ and $y = b + c$ for some integer $-b \leq c \leq a$. Then $xy = (a-c)(b+c) = ab + ca - cb - c^2 = ab - (c(b-a) + c^2)$. Since we assumed that $a \leq b$ and $x \neq a$, we have that c must be positive. Therefore, $c(b-a) + c^2$ is positive and we conclude that $xy < ab$. □

From the above proposition we conclude that the number of terms in the gradient is bounded by $4\hat{d}\lfloor N/2 \rfloor \lceil N/2 \rceil$, with N bounded by $n(n-1)/2$.

6 Principal Component Analysis

A method frequently used to display high dimensional data or reduce the amount of dimensions of a dataset is Principal Component Analysis (PCA). PCA transforms the coordinate system for a dataset, ensuring that the maximal variance of the dataset is explained by the first coordinate, the maximal variance that remains is explained by the second coordinate, and so on. If the variance in higher coordinates is very low we can omit these from the dataset to reduce the number of dimensions. We would like to compare PCA with the results from (metric and ordinal) MDS. This works fine on our simulated data, as we have the points of our strains and antisera in $\mathbb{R}^d$. This is not the case when we're dealing with HI assay data. If we know the relation between the similarities and distances however, we can apply a trick. As mentioned before, we can always fit the n points in $n-1$ dimensions if we only have the distances between these points. We do this in the following manner.

Assuming we have a (random) order of the points. We call the first point x_1 the origin, so $x_1 = (0, 0, \ldots, 0)$. Since x_1 and x_2 together determine a unique line in our high dimensional space. We can pick our basis of this space such that all but the first coordinate of x_2 are equal to 0. We want to ensure that the distance between x_1 and x_2 is equal to the $d_{1,2}$ given by the $\delta_{1,2}$ in the dataset. This leads to the equation $\sqrt{(x_{1,1} - x_{2,1})^2} = \sqrt{(x_{2,1})^2} = d_{1,2}$ which we can use to determine $x_{1,2}$. Note that we have 2 solutions to this equation if $d_{1,2} > 0$. We continue with the fact that x_1, x_2 and x_3 determine a unique plane in our high dimensional space. So we argue that we could pick a basis of our space such that all the coordinates of x_3 besides the first 2 are 0. We have that $\sqrt{(x_{3,1})^2 + (x_{3,2})^2} = d_{1,3}$ and $\sqrt{(x_{3,1} - x_{2,1})^2 + (x_{3,2})^2} = d_{2,3}$ which we can use to determine $x_{3,1}$ and $x_{3,2}$. This argument extends for all points x_i with $i = 1, 2, \ldots, n$. For x_i we get that $x_{i,j} = 0$ when $j \in \{i, i+1, \ldots, n-1\}$. We also get a set of equations $\sqrt{\sum_{k=1}^{i-1} (x_{i,k} - x_{j,k})^2} = d_{i,j}$ for $j \in \{1, 2, \ldots, i-1\}$ which can be used to determine $x_{i,1}, x_{i,2}, \ldots, x_{i,i-1}$. We note that if $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,i-1}, 0, \ldots, 0)$ is a solution to the given set of equations, then $x_i' = (x_{i,1}, x_{i,2}, \ldots, -x_{i,i-1}, 0, \ldots, 0)$ is also a solution. Thus two solutions exist if $x_{i,i-1} \neq 0$, i.e. if we need a new dimension to describe point x_i. Whenever this occurs, we pick one of the two possible solutions. The existence of at least one solution is guaranteed if we are dealing with exact actual distances. The points which generated these distances span a subspace of at most dimension $n-1$ in their original space, these points being a solution to the set of equations.

A problem arises with this method when we have no idea what the relation between the similarities and actual distances is. If the similarities aren't binned we can use $1/\delta_{i,j}$ instead of the $d_{i,j}$'s in the above method to get a result. Though the configuration found this way will obviously not satisfy the exact distances between the original points. It will at least obey the ordering among the distances of the original points, as the similarities obey the inverse ordering. When we are dealing with bins we could use the value assigned to a bin for the above method. This will result in errors, since this value is

merely an indication for the possible real values of the similarities. Usually this value is the mean or upper/lower bound of a bin. An example to what could go wrong is the following. Assume we have 3 points in 2 dimensional space $x_1 = (0, 0), x_2 = (\frac{1}{\sqrt{2}}, 0)$ and $x_3 = (0, \frac{1}{\sqrt{2}})$. The 3 distances then are $d_{1,2} = \frac{1}{\sqrt{2}} = d_{1,3}$ and $d_{2,3} = 1$. Also assume we have 4 bins, with each distance in each bin getting the minimum of the bin assigned as value. The 4 bins are $[0, \frac{1}{3}), [\frac{1}{3}, \frac{2}{3}), [\frac{2}{3}, 1)$ and $[1, \infty)$. So our distances then become $d_{1,2} = d_{1,3} = \frac{1}{3}$ and $d_{2,3} = 1$. It should be immediately clear that this poses a problem, since these distances no longer satisfy the triangle inequality. So we could create a system which simply has no solution.

When looking at binned data, the value each distance (or more generally, each similarity) has is merely an indicator of the possible range of values of this distance (or similarity). In the example above for instance, $d_{1,2} = \frac{1}{3}$ indicates that $d_{1,2} \in [\frac{1}{3}, \frac{2}{3})$. We could incorporate this into our method, by setting that the distances between points should lie between the bounds of their bins. Then trying to find a solution to this set of inequations. We then get a whole lot of roots, containing coordinates, which no longer need to be a specific value but now need to lie between an upper and lower bound, namely the upper and lower bound of the bin our distance lies in. The trouble with efficiently finding a solution to this set of inequations is that we can no longer iteratively solve it. When we have a set of equations to solve as before, we can do them iteratively starting with solving the solutions for our second point, which had only 1 non-zero coordinate. Substituting the solutions we found for previous points greatly reduces the complexity of each equation to solve. Now we have inequations instead of equations, which could be violated if we just iteratively started picking values for our points. Having a computer solve these inequations, e.g. plugging them into mathematica, is infeasible as the number of inequations to satisfy quickly grows with the number of points.

7 The dataset of Smith et al

Smith et al [5] published a paper on the antigenic evolution of influenza, using a large dataset of results from HI assays spanning over 30 years of research. The dataset can be found online, see [7]. The dataset includes 273 strains of influenza and 79 antisera. Resulting in $79 \cdot 273 = 21567$ strain-antiserum distances which could be measured by the HI assays. The dataset online shows a total of 4252 measured values, while the original paper [5] reports 4215 values were measured. I assume the online dataset has been updated slightly since the publication of the article. Of the measured values 3279 are exact values while the remaining 973 are threshold values. When attempting to reconstruct the exact error function used by Smith et al, we failed to recover the same value of the function in the minimum found in the aforementioned paper. This configuration in 2 dimensions was also not the minimum when using the aforementioned dataset and the function $e_{met,thres}$ from (**??**). We continued the minimization from this point onwards and got the configuration which can be seen in figure 3. We note that the picture does not change much. The points seem to move closer to each other, but the cluster structure appears to stay the same.

As was shown in the article above, the error prediction technique used by Smith et al can conclude that 2 dimensions are enough to describe a high dimensional model of antigenic influenza. Thus we wonder what the real dimension is in which antigenic evolution takes place. We could attempt to find the dimension in which the value of our error function is 0 when using the large dataset of Smith et al. Due to measurement errors and the binning in the data however, we cannot hope to find a perfect fit at all.

Instead of attempting to interpret the outcomes of the metric MDS algorithm on the HI assay dataset, we tried to fit the data in several dimensions using the ordinal version of the MDS algorithm as introduced by Lapedes et al [4]. Because the complexity of the ordinal MDS algorithm is rather high, we could not expect results from the algorithm in a reasonable time frame when using the entire dataset. Instead we isolated clusters of points from the dataset manually and used these for the algorithm. Here we'll present results of 2 of these clusters. The first consisting of strains 153-175 and antisera 41-45 in the dataset of Smith et al. The second consisting of strains 222-238 and antisera 48-54 in the dataset of Smith et al. Creating a total of $23 \cdot 5 = 115$, respectively $17 \cdot 7 = 117$, strain-antisera distance which could be measured. Of these 113, respectively 105, were actually measured. We attempted fits in $2, 3, 4, 5$ and 6 dimensions using the error function $e_{ord,thres}$ from (6). Argueably the minimum value of $e_{ord,thres}$ is even harder to interpret than the value of the metric version, as the function does not even need to be 0 when the configuration obeys the ordering given by the similarities (for more details see section 4). Here we present the amount of errors instead, counted using the method explained in section 4 (table 1).

We note that if we only accept a configuration if it makes 0 errors in the ordering

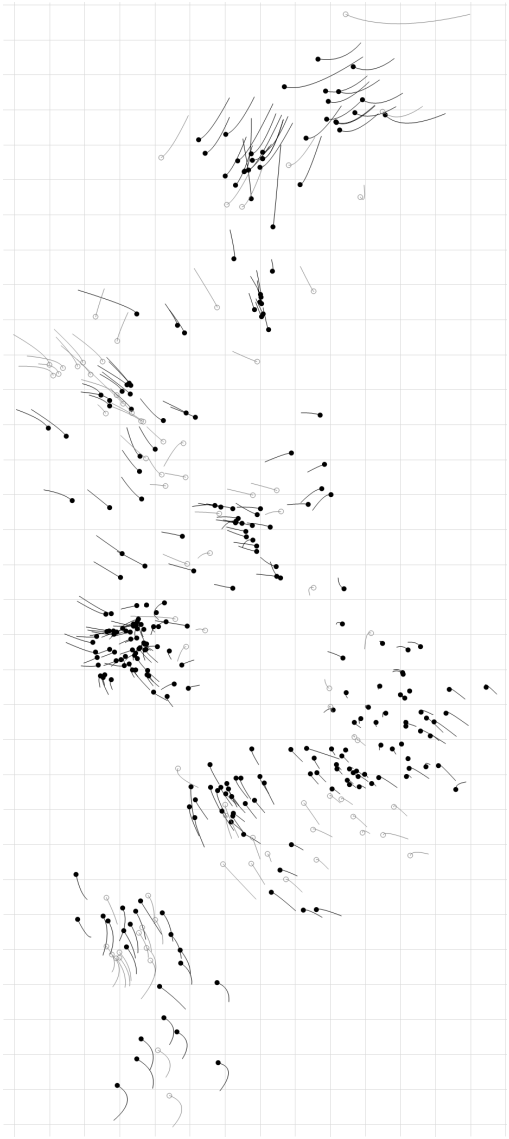

Figure 3: The minimum found from the configuration found by Smith et al using our error function $e_{met,thres}$. The black points indicate the strains and the white points indicate antisera. The lines indicate the change from the configuration found by Smith et al to the configuration we found.

of the points then we need 5 and 6 dimensions respectively to map the clusters of the dataset used in these runs. As before we want to account for errors in the measurements, meaning we don't want to enforce a strict threshold of 0 errors. Instead we construct a different threshold which depends on the dataset. Assume a bin in our dataset contains many similarities and we have another similarity which is measured as lower than the

run	2 dim map	3 dim map	4 dim map	5 dim map	6 dim map
1	868	183	55	0	0
2	2052	1505	701	261	0

Table 1: The amount of errors, when using the method explained in section 4, for the best fit in each dimension using the ordinal MDS algorithm

value of this bin, while it is in fact higher. If we would perform an error counting on the original points, using the dataset we have to obtain the ordering we want to preserve, we would get at least as many errors as lie in this large bin. Therefore, we construct a threshold based on the largest bin in our dataset. The largest bins contain 14 and 7 similarities for the clusters used in run 1 and 2 respectively. Let us assume that for at most 20% of our similarities an error was made during the measurements, so 26 and 21 similarities respectively. We then propose thresholds of $14 * 26 = 364$ and $7 * 21 = 147$ for runs 1 and 2 respectively. We see that the cluster used in run 1 fits in 3 dimensions while the cluster used in run 2 still needs 6 dimensions to be acceptable. We note that the threshold we constructed can be varied depending on the confidence in the given dataset, by varying the percentage of data that is expected to consist of measurement errors.

8 High dimensional model for antigenic evolution

The model used in the article is based on a 15 dimensional random walk. Each point is generated from the previous point p by adding a random change of $X_i^p \sim N(0, \sigma)$ to the coordinates $i = 1, 2, \ldots, 15$ with either $\sigma = 1$ or $\sigma = 25$. Equivalently, each point in generated by adding random a vector X to the previous point. Where X is distributed according to the 15 dimensional multivariate normal distribution with mean the zero vector $\mathbf{0}$ and covariance matrix the identity matrix I or 25 times the identity matrix, i.e. $25I$. We named the step for which the variance is larger large steps in the article, let us justify this name.

The distance between two adjacent points in the random walk is given by $D = \sqrt{\sum_{i=1}^{15} X_i^2}$, which we know to have the chi distribution with 15 degrees of freedom when $X_i \sim N(0, 1)$. In the case of $\sigma = 5$ we can apply a trick. We note that $D = \sqrt{\sum_{i=1}^{15} \frac{X_i^2 \cdot 25}{25}} = 5 \cdot \sqrt{\sum_{i=1}^{15} \frac{X_i^2}{25}}$. We have that $\sqrt{\sum_{i=1}^{15} \frac{X_i}{25}}$ is also chi distributed with 15 degrees of freedom. Therefore our larger distances are given by $5Y$ with Y chi distributed in 15 degrees of freedom, as above. We have that $\mathbb{P}(5Y \leq a) = \mathbb{P}(Y \leq a/5)$, from which we can infer the cumulative distribution function. In figure 4 the cumulative distribution functions of both D and $5Y$ have been plotted.

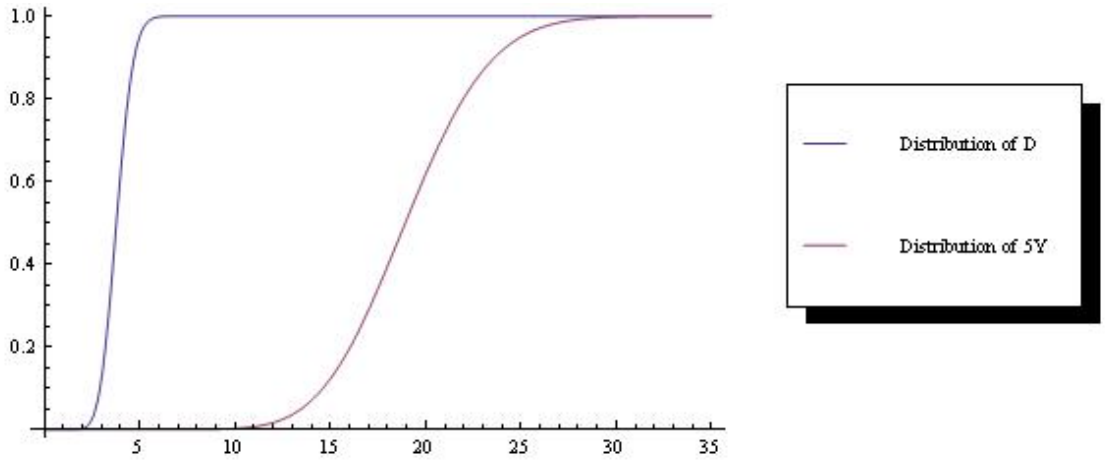

Figure 4: The cumulative distributions of D and $5Y$

As can be seen from the plot, the probability that a large distance, with distribution $5Y$, is smaller than a small distance, with distribution D, is very small. Justifying the notion of smaller and larger distances.

After generating these points and their interpoint distances, we discarded many so our data would resemble the actual data gathered through HI assays. In figure 5 a plot of

the matrix used by Smith et al can be found. This figure depicts the bias in the data gathered towards strains and antisera which lie close in time.

In our model we discarded larger distances for the reason that those cannot be accurately measured by the HI assays [8]. The points we used in our model are generated using an iterative process, which means there is a natural time ordering between the points. Because our random walk takes place in 15 dimensions, we can assume the process never returns to the previous points generated because we have only a limited number of points. In the sense that with 15 coordinates changing in each step, the chance that each of them changes back to (nearly) the previous value is negligible. Thus we expect there to be an increasing distance between the points which lie further apart chronologically. These are exactly the distances we'll throw away in our model.

9 Simulations on the effect of discarding large distances

We wish to get a better understanding of what the effect of discarding large distances is on the results of the metric MDS algorithm. For this purpose we did several simulations to see the effects on the results in different settings. Here we'll present these results case by case with an explanation of what we modeled exactly and an interpretation of the results.

9.1 Only discarding the large distances

Once more we use a random walk, as we did in the model in the article, to generate 50 points in dimensions $5, 10, 15, 20$ and 25. We do not make a distinction between large and small steps, the change per coordinate for each step is a standard normally distributed random variable. We do not add points for the antisera either, to keep the computation time of the algorithm somewhat in check. Of all the distances between these 50 points generated we discarded the largest 75%. We attempted to fit them in dimensions 2 through d, with d the dimension in which the data was generated. We calculated the average error between the actual distances which were discarded and the distances in the resulting map, the prediction error from the article. We repeated this for 27 runs. The results of the runs combined can be seen for the different dimensions in figures 6, 7, 8, 9 and 10.

The figures show that the prediction error does not change much when we increase the target dimension of the algorithm. This holds true for all original dimensions we used, the dimensions in which we generated the random walk. We conclude that the notion of prediction error cannot be used to accurately estimate the original dimension of data used by the MDS algorithm when large distances are missing. Which is exactly the case in data found through HI assays. We note that the random walk used to generate the data presented here only generates points for strains and not for antisera. In the case of HI assay data we only have antisera-strain similarities. If we would designate some points as antisera in the random walk we generated, we'd only have information about the distances between those points and all the other points not designated as antisera. In a way, the model we use here has more information as we have the distances between all the points, before we started discarding.

9.2 Varying the number of distances discarded

We use the same random walk as above, but we vary the number of distances discarded in an attempt to see if we can produce better results should we be able to measure more of the small similarities/distance. Instead of discarding the largest 75% of the distances we removed the largest 50% and 25% instead. In both cases we only did 15 runs to limit the time required for the simulation. Results when the largest 50% was discarded can be found in figures 11, 12 and 13. Results when the largest 25% was discarded can be

found in figures 14, 15 and 16.

We see an improvement in the prediction error when we go from target dimension 2 to 3, but not when we increase the target dimension even further. We conclude that even if we would greatly improve the range of similarities we can measure for use in the metric MDS algorithm, we still cannot accurately estimate the dimension of the underlying shape space using the prediction error.

9.3 Increasing the size of the random walk

To see the long term effect of discarding large distances, because they either cannot be measured by the HI assay or were tested in two different HI assays, we increased the amount of points generated in the previously used random walk in 15 dimensions to 200. For completeness we also associated antisera to strains with a probability of 1/3.5, with the antiserum being the exact same point as the serum it is associated with. We then discarded the largest 50% of the distances and used the metric MDS algorithm to fit the strains and antisera in 2 dimensions (figure 17).

The strains loop back onto each other, while the distance between two points increases as they lie further apart on the random walk used to generated the actual distances between the points. This is a curious effect caused by the high dimension used to generate the data and the discarding of the larger distances.

10 The limit case $\mathbb{R}^\infty$

We are interested in the behavior of our algorithms in the case that we take our original dimension to be very high, so that it approaches the limit case $\mathbb{R}^\infty$. We start by considering the random walk, in which we only make changes to our coordinates according to the standard normal distribution. In the case of a fixed, finite dimension d the distance between two consecutive points is $D = \sqrt{\sum_{i=1}^{d} X_i^2}$. We know that the random variable D has the chi distribution in d degrees of freedom. The first question at hand is, what happens with this distribution as d becomes large. The mean and variance of a chi distribution in d degrees of freedom are given by:

$$\mu_d = \sqrt{2}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \tag{12}$$

$$\sigma_d^2 = d - \mu_d^2 \tag{13}$$

When we consider the limit of these equations as d goes to ∞ we get the following limits, which can be calculated using mathematica:

$$\lim_{d\to\infty} \mu_d = \lim_{d\to\infty} \sqrt{2}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} = \infty$$

$$\lim_{d\to\infty} \sigma_d^2 = \lim_{d\to\infty} d - \left(\sqrt{2}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\right)^2 = \frac{1}{2}$$

We see that the mean of the distance goes to infinity as we increase the dimension. Since the step we take in each dimension is independent of the other dimensions and has a standard normal distribution, the step in each direction is almost surely positive. So each coordinate makes an almost surely positive contribution to the distance. So the total distances has to increase to infinity as the amount of dimensions does. We note, since the limit is finite, that the variance is bounded when d becomes large.

This only gives us the distribution of the distances between two points that are actually adjacent in the random walk. The question is what happens if they are not. Assume we have 2 points n and m on the random walk with m k steps after n. In that case m_l, the l-th coordinate of m is n_l, the l-th coordinate n, plus k independent variables X_i with $X_i \sim N(0,1)$ for $i = 1, 2, \ldots, k$. The sum $X^{(k)}$ of the X_i's is again a normal variable with mean 0 and variance k. So we can conclude that the distribution of the difference between the l-th coordinate of m and n is equal to $X^{(k)}$. Note that the distribution of $X^{(k)}$ is independent of the coordinate we choose to evaluate.

Now we would like to know how the distance between n and m is distributed. We know that this distance equals $\sqrt{\sum_{i=1}^{d} (X^{(k)})^2}$. Note that, since $X^{(k)} \sim N(0,k)$, we have that $\frac{X^{(k)}}{\sqrt{(k)}} \sim N(0,1)$. Applying the same trick as in section 9 we get the distribution of the distance:

$$\sqrt{\sum_{i=1}^{d}(X^{(k)})^2} = \sqrt{\sum_{i=1}^{d}\left(\sqrt{k}\frac{X^{(k)}}{\sqrt{k}}\right)^2}$$
$$= \sqrt{k}\sqrt{\sum_{i=1}^{d}\left(\frac{X^{(k)}}{\sqrt{k}}\right)^2}$$

Where we know that $\frac{X^{(k)}}{\sqrt{k}} \sim N(0,1)$. So the second square root is a random variable with a chi-distribution in d degrees of freedom.

It is not immediately clear that this limit case has nice properties. For this reason we apply a trick. We rescale our coordinates by a factor $1/\mu_d$. We can do this because rescaling doesn't intrinsically change the picture we get, which is all we're interested in. Due to this rescaling factor the distance between two points is rescaled by the same factor $1/\mu_d$. We see that the mean and variance of our distribution function for the distance between two adjacent points in the random walk go to 1 and 0 respectively. We conclude that this particular limit case is a deterministic process, with the distance between two adjacent coordinates being 1. From the distribution of the distance between the n-th and m-th point of our random walk we can infer that this is now a step of fixed distance $\sqrt{|n-m|}$.

We are interested in the outcome of our algorithms when we use these values as input. Assume we generate n points according to the above relation. Let us first consider the ordinal algorithm. Since we're dealing with ties here, we need the version of the algorithm which incorporates them. This algorithm will always find a perfect fit in dimension 1. Because no matter how many points we generate, $x_1 = 0, x_2 = 1, \ldots, x_n = n-1$ is always a configuration which obeys the ordering of the distances. Note also that multiplication or addition by any real number does not violate this fact, so any configuration generated from the given one by such manipulations is also a solution for the algorithm in dimension 1, which does not make errors in the ordering.

Let us now consider the metric MDS algorithm. If we consider target dimension 1 ,we can prove the following:

Proposition 2. *The points $x_0 = 0$ and $x_i = \frac{1}{n}\sum_{k=1}^{i-1}\sqrt{k} + \sqrt{n-k}$ for $i = 1, 2, \ldots, n$ are a local minimum of the metric MDS error function e_{metric} (2) w.r.t. the distances given above between n points.*

Proof: We'll show that the gradient is equal to 0 for the points and distances given in the proposition. For this we obviously need the gradient of the metric MDS error function. We have computed this before, it is given by (8). First we note that in dimension 1, each point only has one coordinate. So the distances between two points x_i and x_j is the absolute value of the difference between the two points. Consider coordinate k of the gradient:

$$\frac{\partial e_{metric}}{\partial x_{k,1}} = -2 \sum_{j=1, j\neq k}^{n} (D_{k,j} - |x_k - x_j|) \frac{x_k - x_j}{|x_k - x_j|}$$

We have set $D_{k,j} = \sqrt{k-j}$ for $k > j$ and $D_{k,j} = \sqrt{j-k}$ for $k < j$. We also know that $|x_k - x_j| = x_k - x_j$ for $k > j$ and $|x_k - x_j| = -(x_k - x_j)$ for $k < j$ since our points obey $x_1 < x_2 < \ldots < x_n$. We can use this to simplify the above, then we can use the expressions for the points from the proposition. We get:

$$\begin{aligned}
&-2 \sum_{j=1, j\neq k}^{n} (D_{k,j} - |x_k - x_j|) \frac{x_k - x_j}{|x_k - x_j|} \\
=&-2 \left(\sum_{j=1}^{k-1} (D_{k,j} - x_k + x_j) \frac{x_k - x_j}{x_k - x_j} + \sum_{j=k+1}^{n} (D_{k,j} + x_k - x_j) \frac{x_k - x_j}{-(x_k - x_j)} \right) \\
=&-2 \left(\sum_{j=1}^{k-1} (D_{k,j} - x_k + x_j) - \sum_{j=k+1}^{n} (D_{k,j} + x_k - x_j) \right) \\
=&-2 \left(\sum_{j=1}^{k-1} (\sqrt{k-j} - \frac{1}{n} \left(\sum_{l=1}^{k-1} \sqrt{l} + \sqrt{n-l} + \sum_{l=1}^{j-1} \sqrt{l} + \sqrt{n-l}) \right) \right) \\
&+2 \left(\sum_{j=k+1}^{n} (\sqrt{j-k} + \frac{1}{n} \left(\sum_{l=1}^{k-1} \sqrt{l} + \sqrt{n-l} - \sum_{l=1}^{j-1} \sqrt{l} + \sqrt{n-l}) \right) \right) \\
=&-2 \left(\sum_{j=1}^{k-1} \sqrt{k-j} - \sum_{j=k+1}^{n} \sqrt{j-k} - \frac{1}{n} \left(\sum_{j=1}^{k-1} \sum_{l=j}^{k-1} \sqrt{l} + \sqrt{n-l} + \sum_{j=k+1}^{n} \sum_{l=k}^{j-1} \sqrt{l} + \sqrt{n-l} \right) \right)
\end{aligned}$$

This expression seems somewhat cumbersome, but it can be simplified using the following identities:

$$\begin{aligned}
\sum_{j=1}^{k-1} \sum_{l=j}^{k-1} \sqrt{l} + \sqrt{n-l} =& \sqrt{1} + \sqrt{n-1} + 2\sqrt{2} + 2\sqrt{n-2} + \ldots \\
&+ (k-1)\sqrt{k-1} + (k-1)\sqrt{n-(k-1)} \\
=& \sum_{l=1}^{k-1} l(\sqrt{l} + \sqrt{n-l}) \\
\sum_{j=k+1}^{n} \sum_{l=k}^{j-1} \sqrt{l} + \sqrt{n-l} =& (n-k)\sqrt{n-(n-k)} + (n-k)\sqrt{n-k}
\end{aligned}$$

$$
\begin{aligned}
&+ (n-(k+1))\sqrt{n-(n-(k+1))} + (n-(k+1))\sqrt{n-(k+1)} \\
&+ \ldots + \sqrt{n-1} + \sqrt{1} \\
&= \sum_{l=1}^{n-k} l(\sqrt{l} + \sqrt{n-l}) \\
\sum_{j=k+1}^{n} \sqrt{j-k} &= \sqrt{1} + \sqrt{2} + \ldots + \sqrt{n-k} \\
&= \sum_{l=1}^{n-k} \sqrt{l}
\end{aligned}
$$

Using these identities, the expression we had for the k-th coordinate of our gradient becomes:

$$
-2\left(\sum_{l=1}^{k-1}\sqrt{l} - \sum_{l=1}^{n-k}\sqrt{l} + \frac{1}{n}\left(\sum_{l=1}^{n-k} l(\sqrt{l}+\sqrt{n-l}) - \sum_{l=1}^{k-1} l(\sqrt{l}-\sqrt{n-l})\right)\right)
$$

Now we can distinguish between the cases $n-k \geq k$ and $n-k < k$. First assume $n-k \geq k$. Then the above simplifies to:

$$
\begin{aligned}
&-2\left(-\sum_{l=k}^{n-k}\sqrt{l} + \frac{1}{n}\sum_{l=k}^{n-k} l(\sqrt{l}+\sqrt{n-l})\right) \\
&= -2\left(-\sum_{l=k}^{n-k}\sqrt{l} + \frac{1}{n}\sum_{l=k}^{n-k} n\sqrt{l}\right) \\
&= 0
\end{aligned}
$$

Where we used:

$$
\begin{aligned}
\sum_{l=k}^{n-k} l(\sqrt{l}+\sqrt{n-l}) &= k(\sqrt{k}+\sqrt{n-k}) + (k+1)(\sqrt{k+1}+\sqrt{n-(k+1)}) + \ldots \\
&\quad + (n-k)(\sqrt{n-k}+\sqrt{n-(n-k)}) \\
&= \sum_{l=k}^{n-k} n\sqrt{l}
\end{aligned}
$$

The case $n-k < k$ has the same equations, except that the summation limits are switched. Therefor this case also satisfies the desired property. □

When we increase the target dimension, the formula for each coordinate of the gradient because more complex. A general solution to these equations was not found. Nonetheless we can still present a picture of the best fit found in 2 dimensions by the algorithm of 100 points with inter-point distances as described above (figure 18).

11 To do list

During the time I took to do the research presented here, I explored a great many options for the ordinal and metric MDS algorithms. A great deal of things were also left untouched or only briefly touched. Here is a short summary of topics and problems I would've liked to look into as well. Anyone who continues with the research presented here may keep this list in mind.

1. It should be clear that the right censoring, the discarding of large distances, greatly affects the outcomes of the MDS algorithm. As postulated in the article, future studies should focus on ways of avoiding this problem through matrix completion algorithms or similar techniques [9].

2. As explained in section 6, under certain assumptions the number of terms in the various error functions and gradeints of the error functions are polynomially bounded. Unfortunately this does not guarantee a low running time, which is not the case in pratice. It should be wise to try and improve the algorithm's running time for future studies.

3. The number of simulations done for the results presented in section 10 was limited due to time constraints and the aforementioned high computation time of the algorithms. Though we don't expect the results to change when we increase the number of runs, it should still be done for completeness.

4. The metric MDS algorithm as presented here attempts to fit distances to an exact value gained through HI assays. But the HI assay suffers from interval censoring. It might be possible to alter the error function used in the algorithms so it does not penalize the distance between points in a configuration if it lies within a range of values, instead of when it is exactly equal to the target distance. Though such a modification could result in insurmountable computation times.

5. The linear relation between the HI assay data and the actual distances in shape space [4, 5] was found using the ordinal MDS algorithm, which has a very long running time on larger datasets. Nevertheless I advice that the relation is checked in the case of a larger dataset, possibly by imputing a subset of the dataset.

Acknowledgment

I would like to thank Martin Bootsma for being my supervisor from the university of Utrecht for this thesis and introducing me at the RIVM. I would also like to thank all the people at the RIVM who I've been working with during my internship there for contributing to such a nice environment to work in. Special thanks go towards Rolf Ypma and Tjibbe Donker, my supervisors from the RIVM. They were always supportive and available for any questions. I would also like to thank Kees van den Wijngaard and Liesbeth Mollema of the RIVM for allowing me to use their computers during the weekends for simulations.

In general, I would also like to thank the university of Utrecht and specifically the staff of the mathematics department. The courses I was able to take in the past five years really opened my eyes to both the beauties and difficulties in the vast amount of topics in mathematics. The teachers were always very helpful and contributed to a nice environment to study in, which helped further spark my interests for various topics in mathematics. But maybe equally important were my fellow students during the past five years whom contributed to this environment through their help and friendship.

Finally I would like to thank my parents and sister for the continued support through the past 5 years. You never stopped believing in me.

References

[1] WHO, "Influenza (Seasonal)." `http://www.who.int/mediacentre/factsheets/fs211/en/`, 2009.

[2] A. Perelson and G. Oster, "Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination," *Journal of theoretical biology*, vol. 81, no. 4, pp. 645–70, 1979.

[3] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis," *Psychometrika*, vol. 27, no. 1, pp. 1–27, 1964.

[4] A. S. Lapedes and R. Farber, "The geometry of shape space: application to influenza," *Journal of theoretical biology*, vol. 212, pp. 57–69, Sept. 2001.

[5] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. a. M. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *Science (New York, N.Y.)*, vol. 305, pp. 371–6, July 2004.

[6] G. Hirst, "Studies of antigenic differences among strains of influenza A by means of red cell agglutination," *The Journal of Experimental Medicine*, no. 10, 1943.

[7] D. J. Smith, "Antigenic Cartography." `http://www.antigenic-cartography.org`, 2007.

[8] G. Hirst, "The quantitative determination of influenza virus and antibodies by means of red cell agglutination," *The Journal of experimental medicine*, pp. 49–64, 1942.

[9] Z. Cai, T. Zhang, and X.-F. Wan, "A computational framework for influenza antigenic cartography," *PLoS computational biology*, vol. 6, p. e1000949, Jan. 2010.

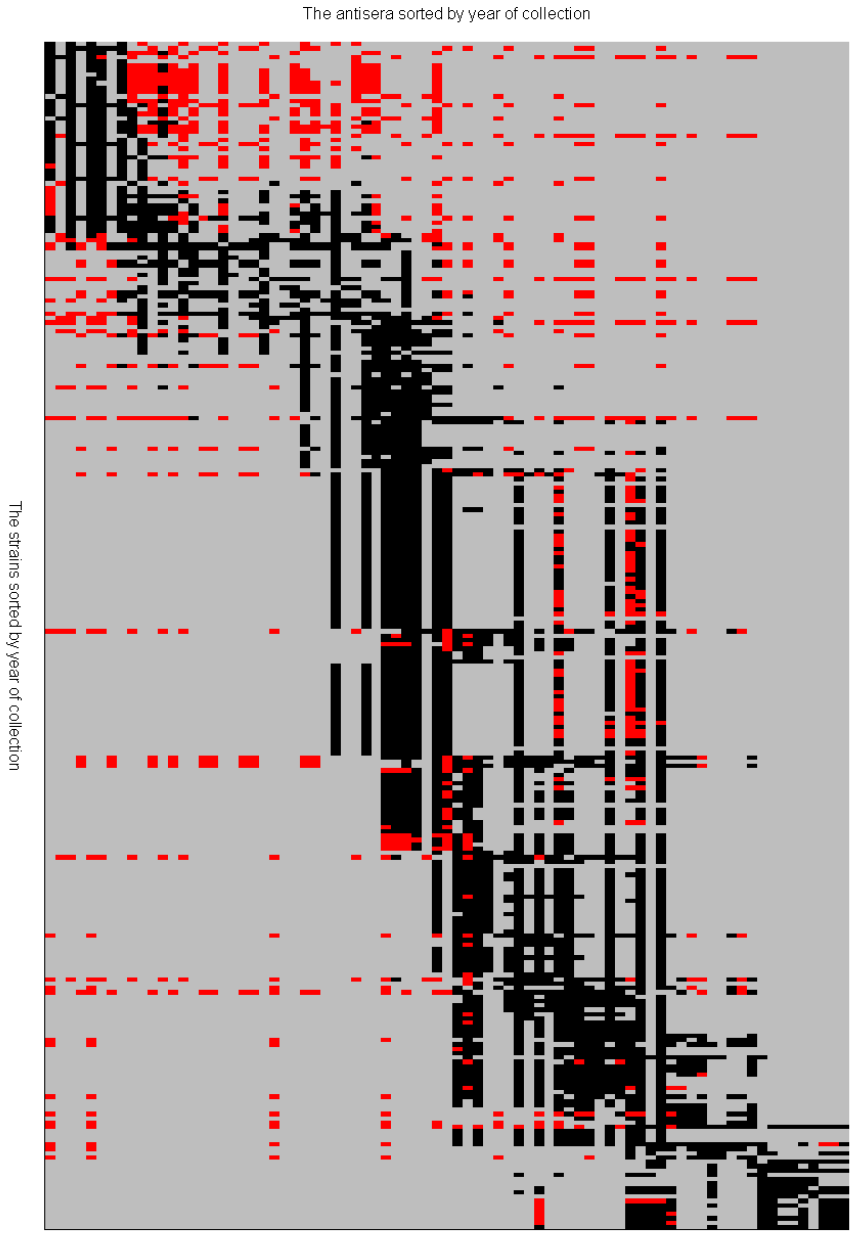

Figure 5: The dataset used by Smith et al. On the horizontal axis we have the antisera and on the vertical axis we have the strains. The strains and antisera have been resorted according their year of creation/isolation. A black square marks an exact value for the strain/antiserum combination. Red squares mark threshold values and grey squares indicate a missing value

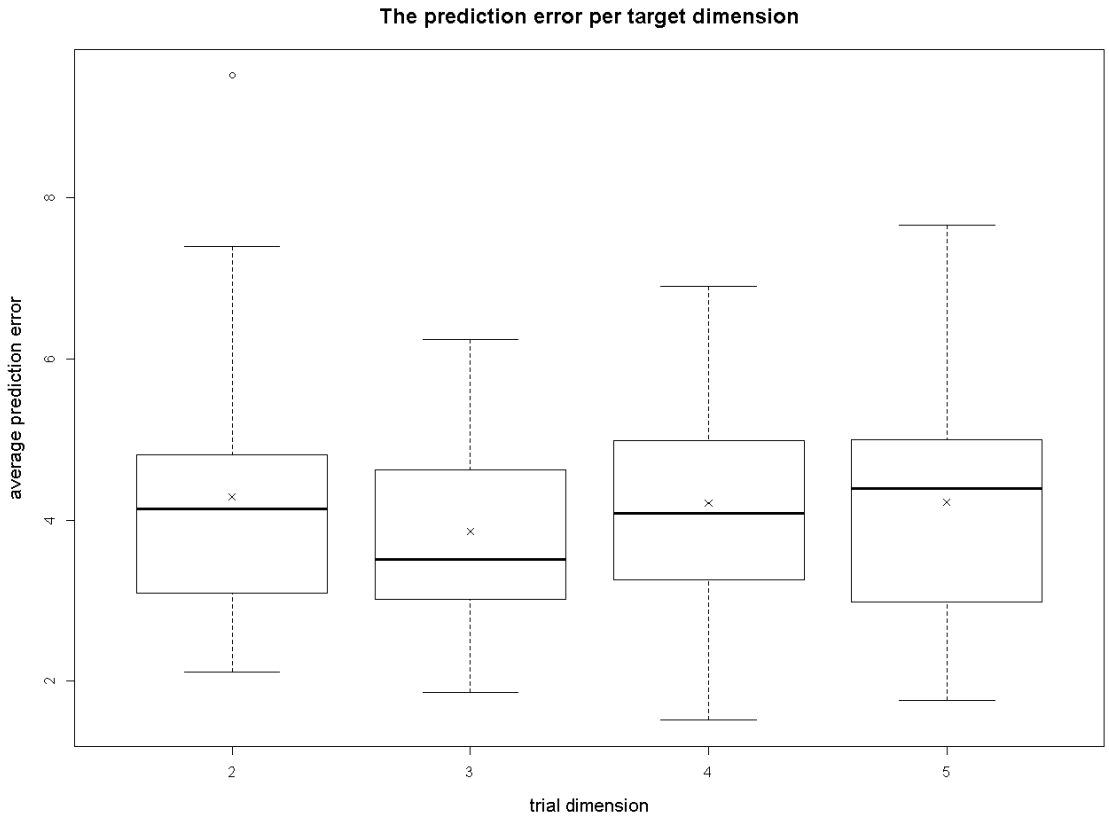

Figure 6: Boxplot of the prediction error of the random walks in dimension 5 which generated 50 points with the largest 75% of the distances discarded. The crosses indicate the average of all the runs per dimension.

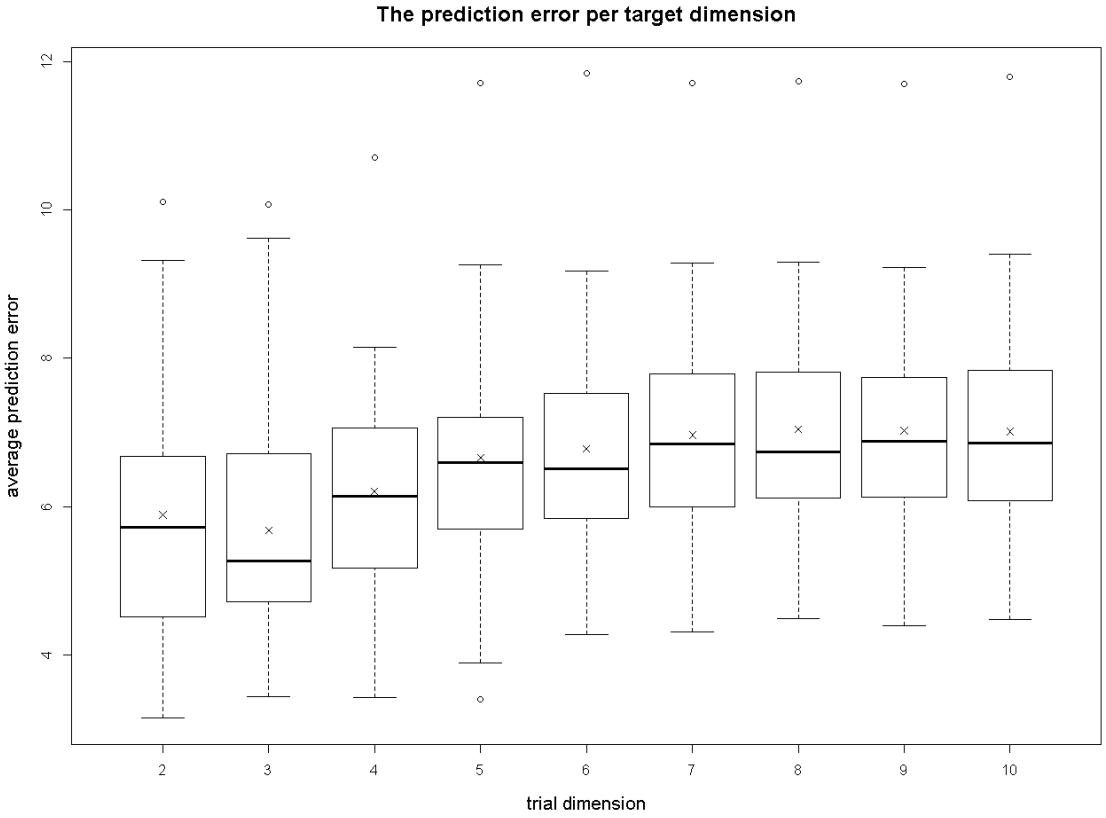

Figure 7: Boxplot of the prediction error of the random walks in dimension 10 which generated 50 points with the largest 75% of the distances discarded. The crosses indicate the average of all the runs per dimension.

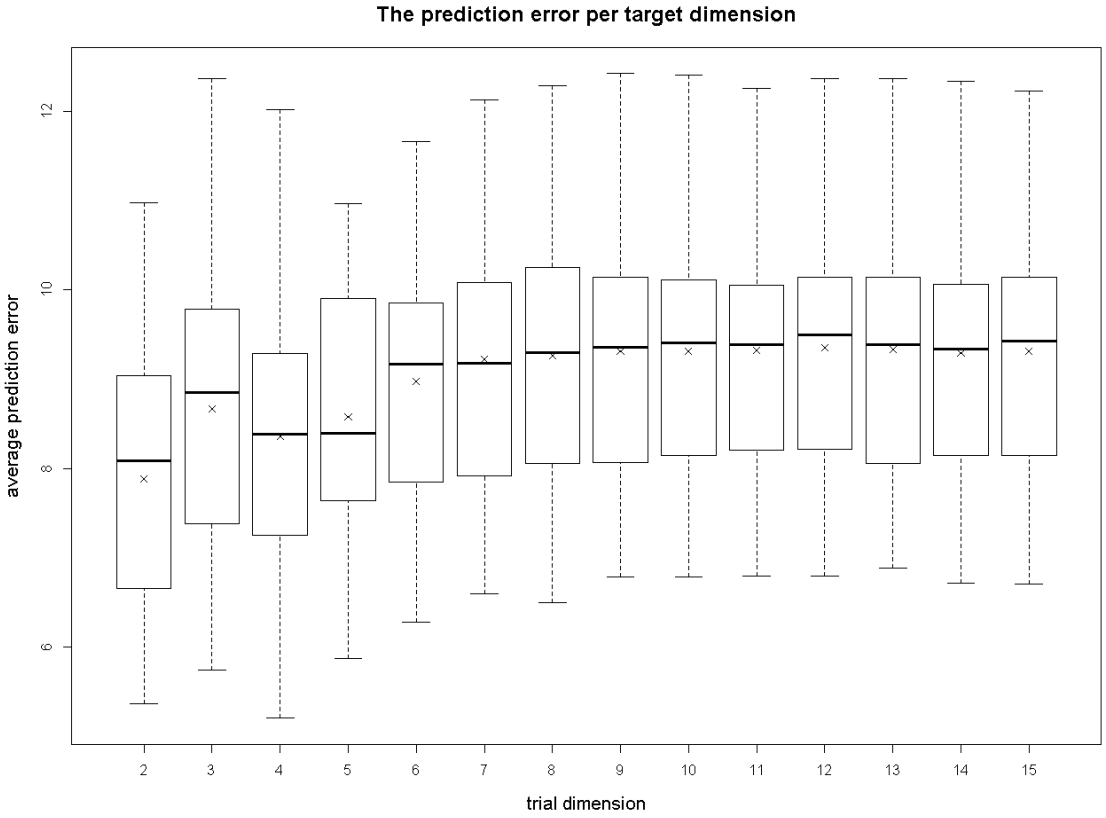

Figure 8: Boxplot of the prediction error of the random walks in dimension 15 which generated 50 points with the largest 75% of the distances discarded. The crosses indicate the average of all the runs per dimension.

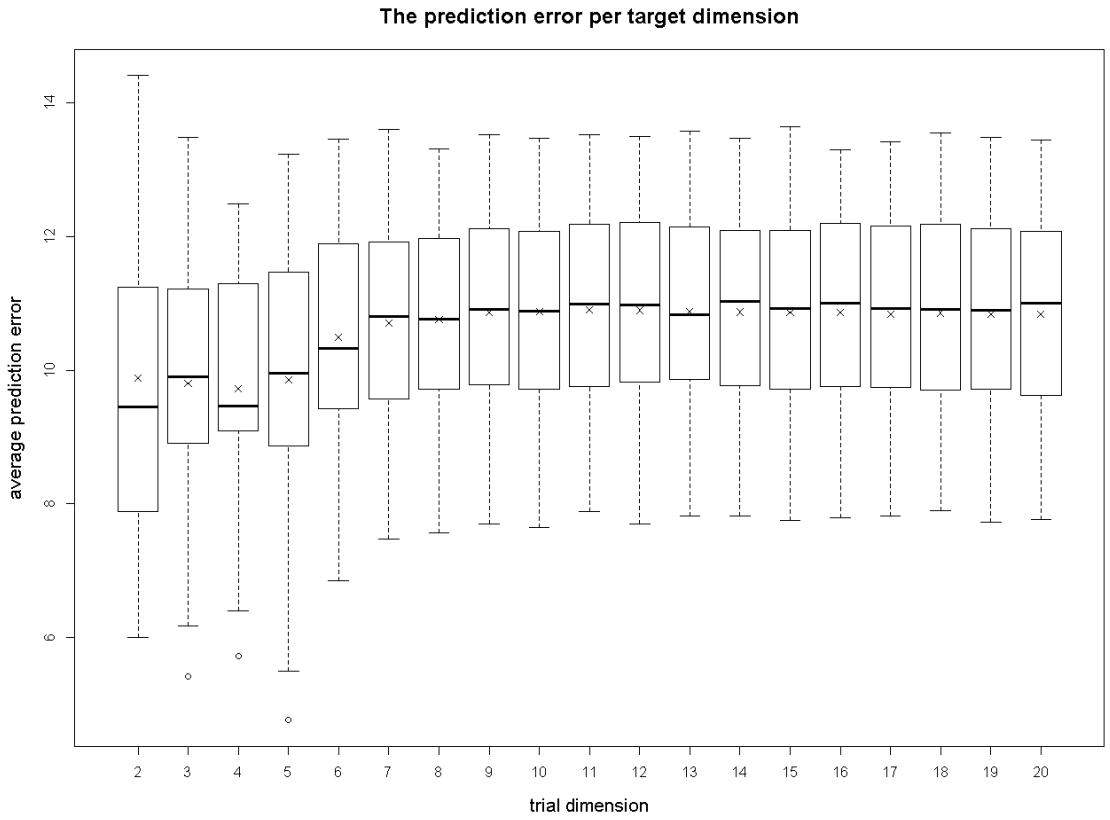

Figure 9: Boxplot of the prediction error of the random walks in dimension 20 which generated 50 points with the largest 75% of the distances discarded. The crosses indicate the average of all the runs per dimension.

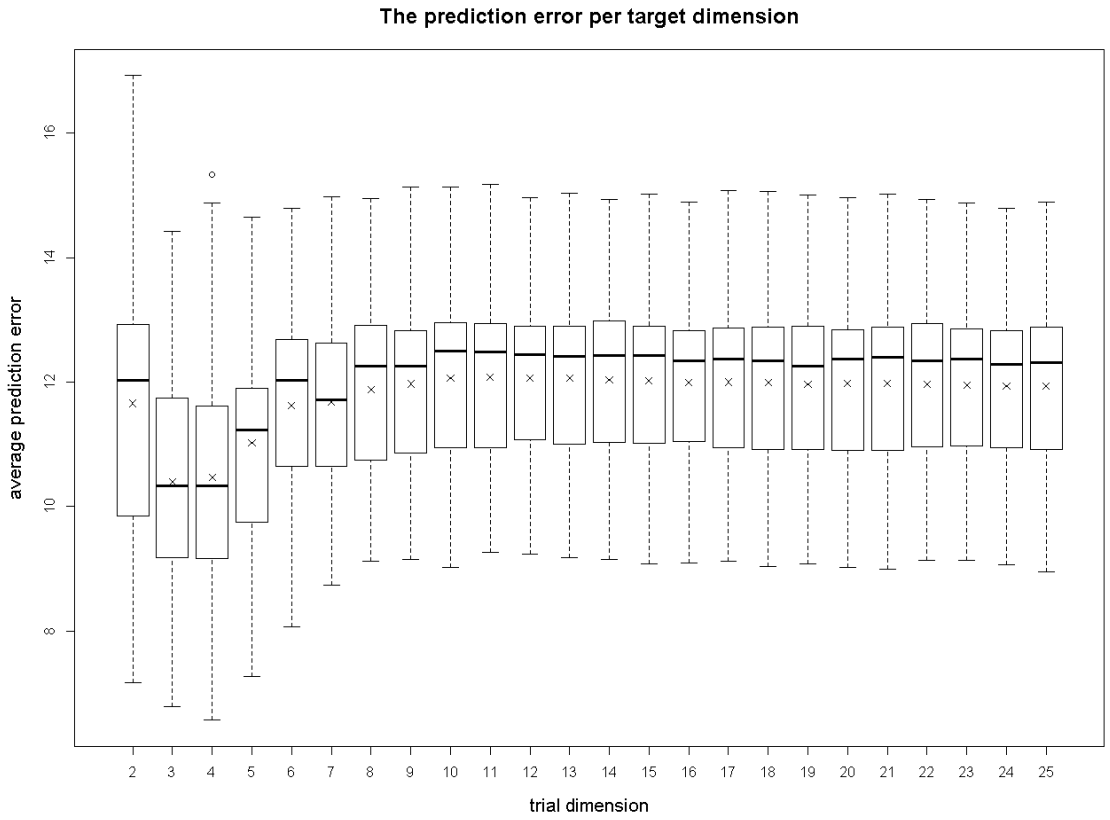

Figure 10: Boxplot of the prediction error of the random walks in dimension 25 which generated 50 points with the largest 75% of the distances discarded. The crosses indicate the average of all the runs per dimension.

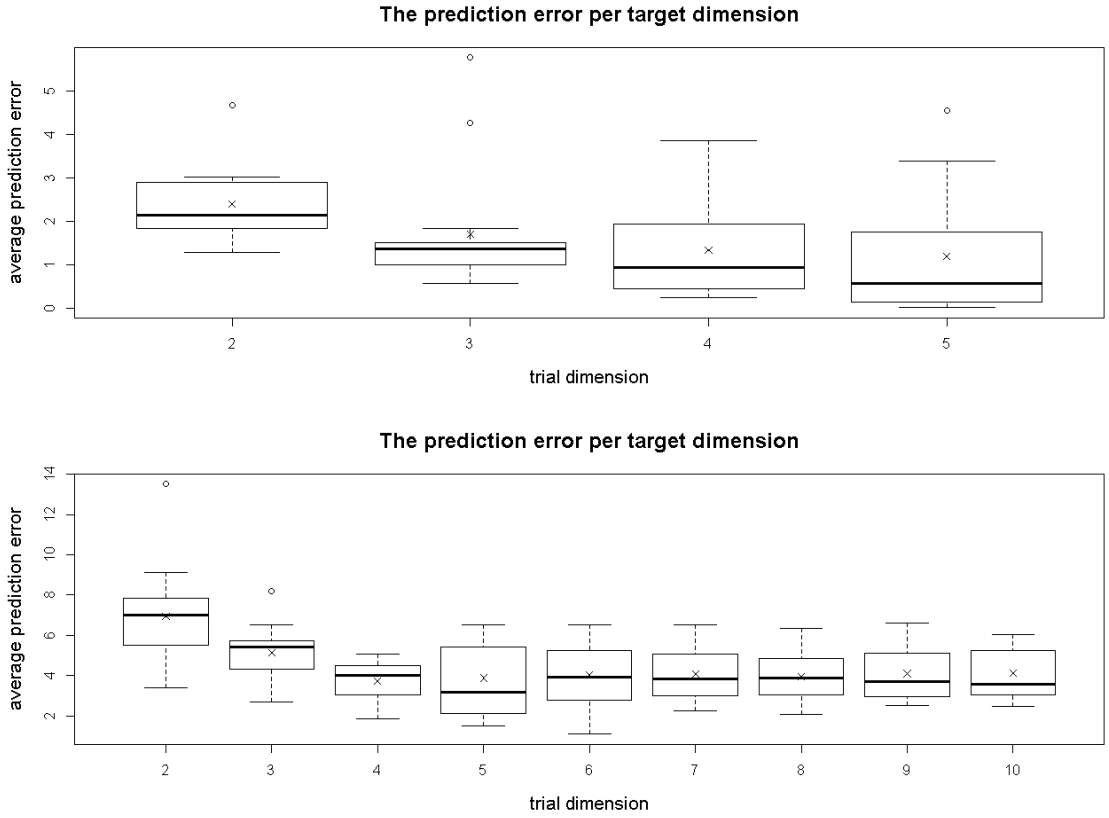

Figure 11: Boxplots of the prediction error of the random walks in dimensions 5 (above) and 10 (below) which generated 50 points with the largest 50% of the distances discarded. The crosses indicate the average of all the runs per dimension.

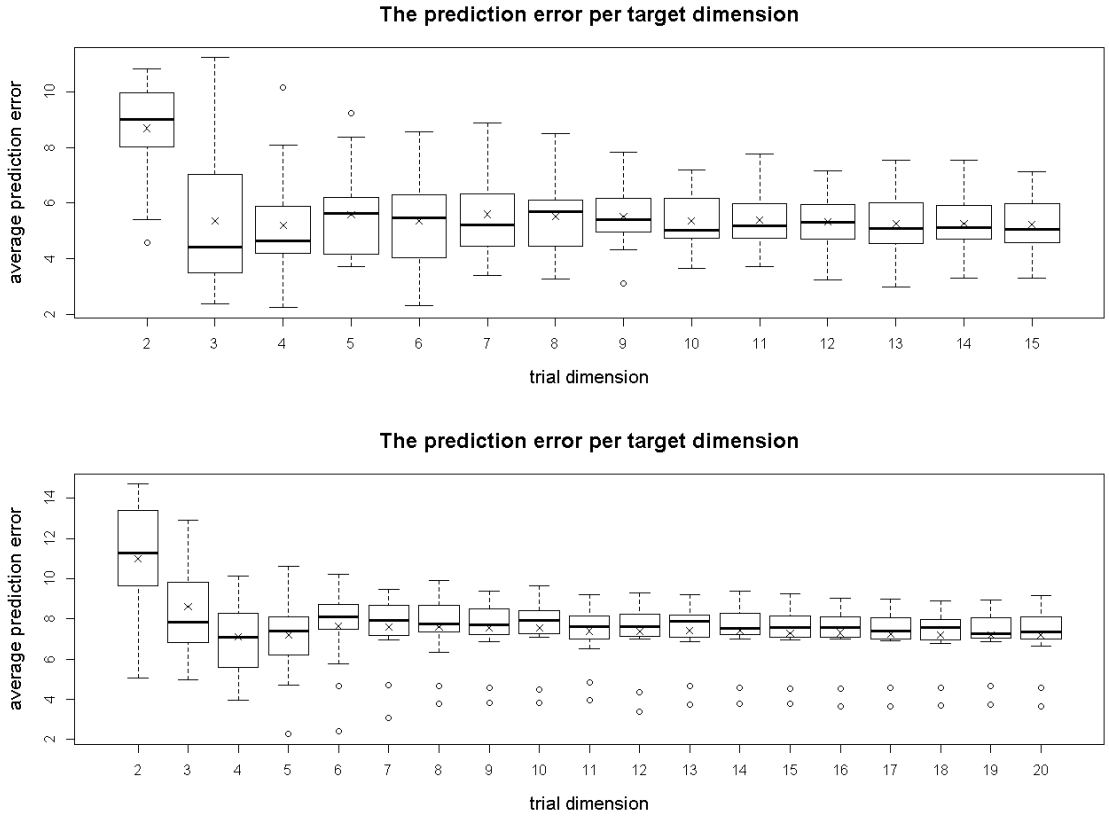

Figure 12: Boxplots of the prediction error of the random walks in dimensions 15 (above) and 20 (below) which generated 50 points with the largest 50% of the distances discarded. The crosses indicate the average of all the runs per dimension.

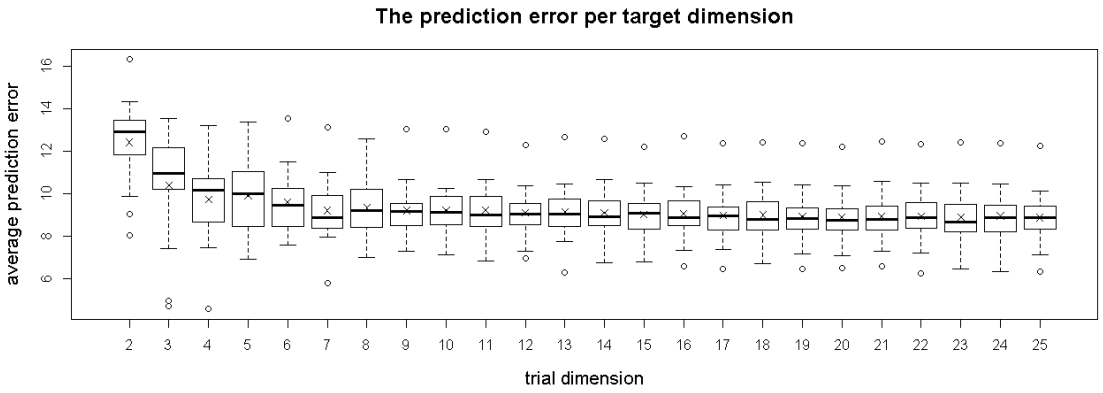

Figure 13: Boxplot of the prediction error of the random walks in dimension 25 which generated 50 points with the largest 50% of the distances discarded. The crosses indicate the average of all the runs per dimension.

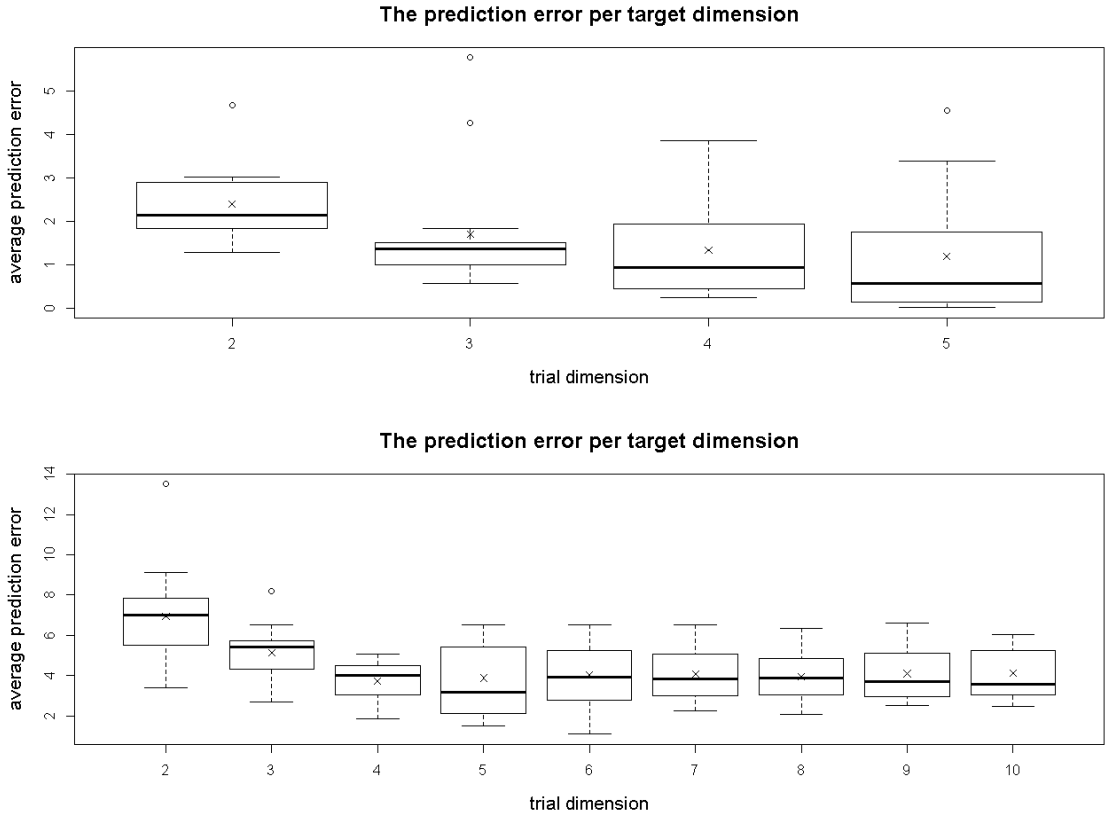

Figure 14: Boxplots of the prediction error of the random walks in dimensions 5 (above) and 10 (below) which generated 50 points with the largest 25% of the distances discarded. The crosses indicate the average of all the runs per dimension.

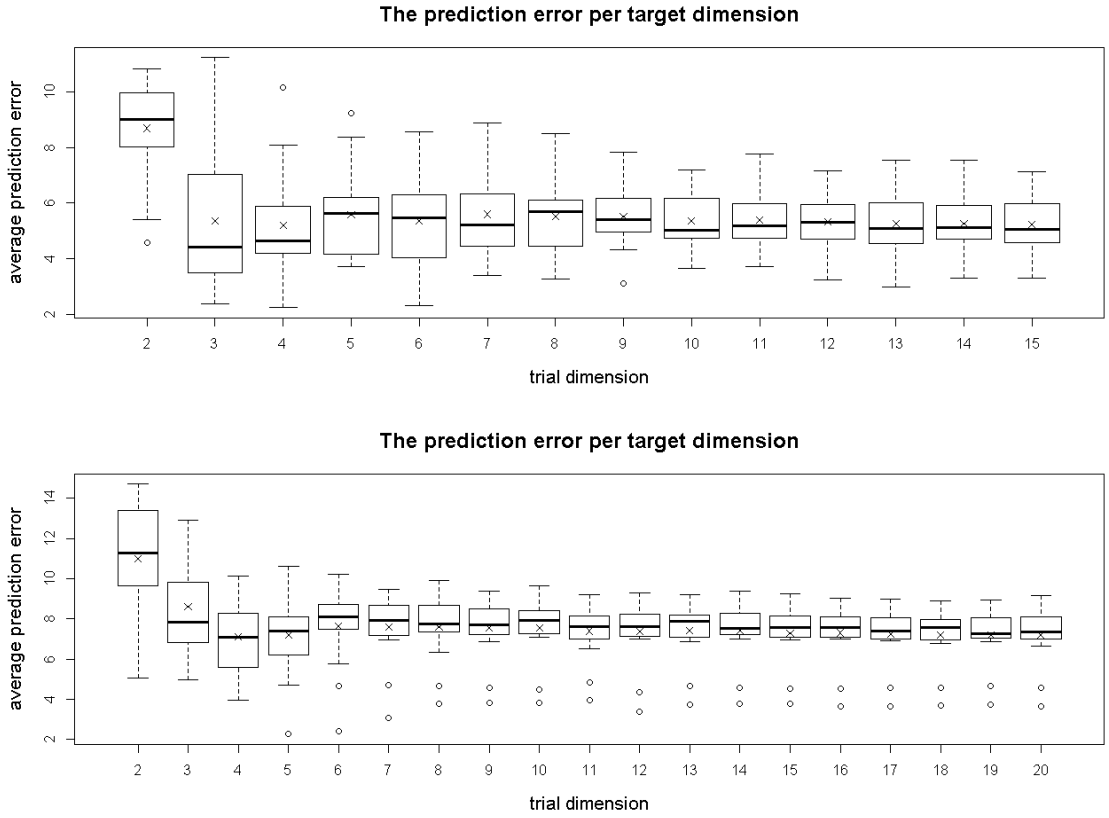

Figure 15: Boxplots of the prediction error of the random walks in dimensions 15 (above) and 20 (below) which generated 50 points with the largest 25% of the distances discarded. The crosses indicate the average of all the runs per dimension.

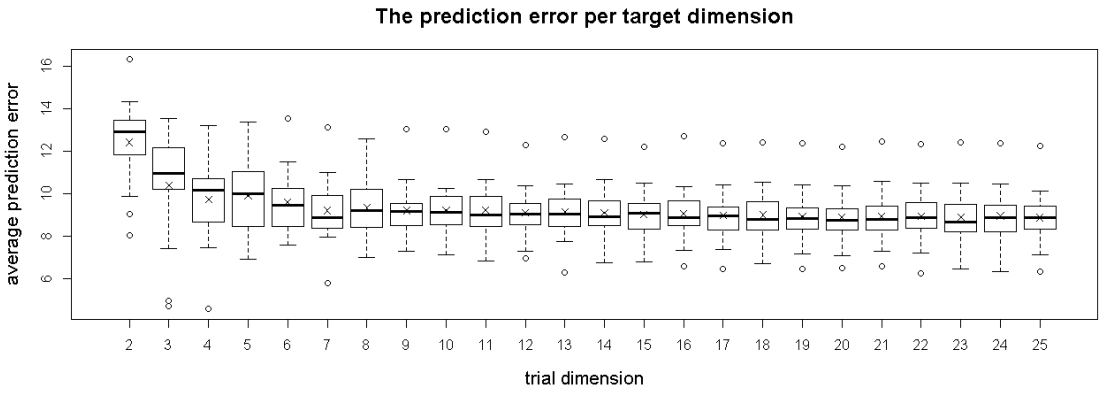

Figure 16: Boxplot of the prediction error of the random walks in dimension 25 which generated 50 points with the largest 25% of the distances discarded. The crosses indicate the average of all the runs per dimension.

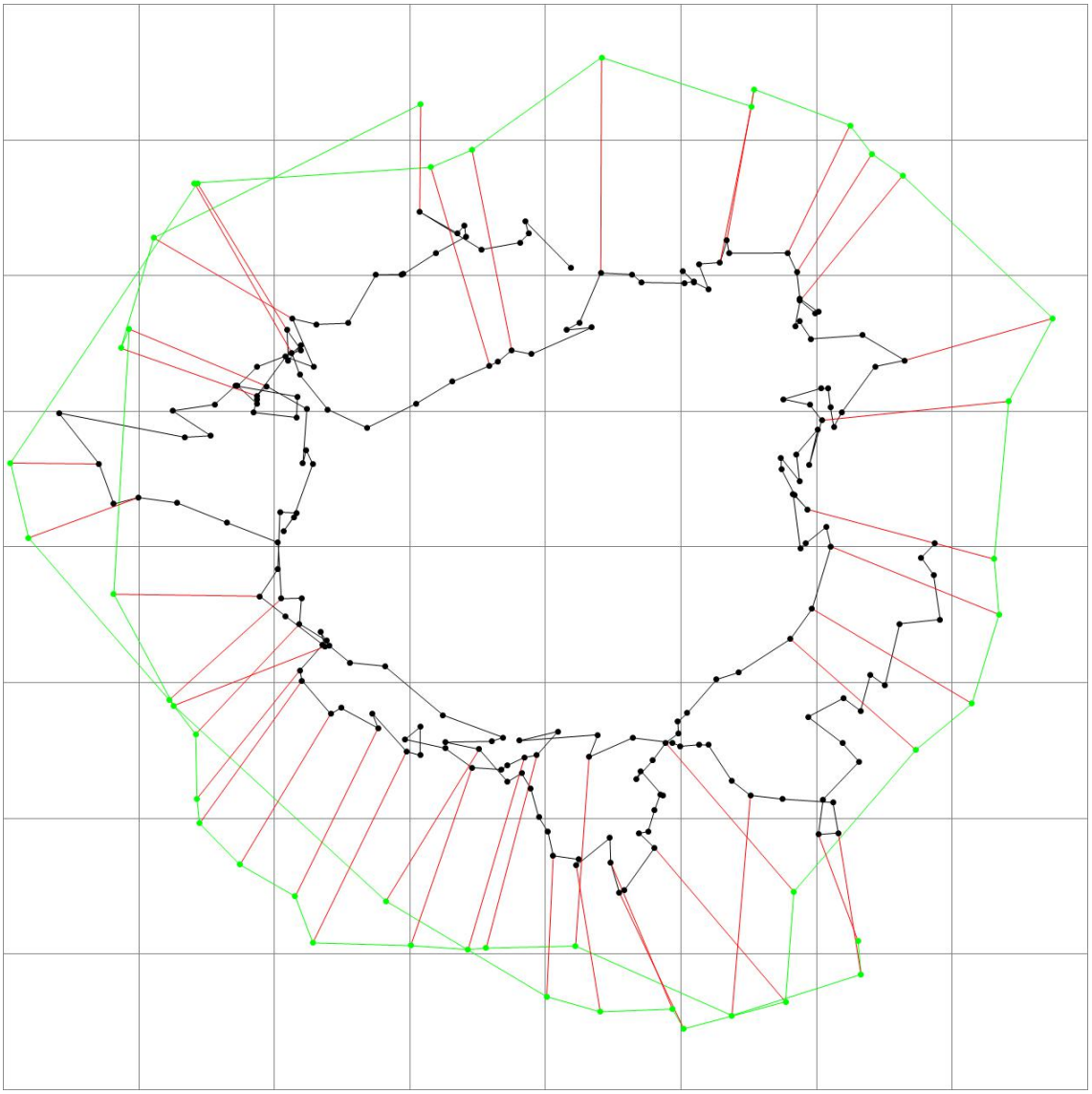

Figure 17: A random walk in 15 dimensions creating 200 strains and associated antisera fit into 2 dimensions using the metric MDS algorithm. The black points are the strains and the green points the antisera. The red lines indicate that the actual distance between the strain and antiserum is 0. The points start to overlap, even though the actual distance between points further apart on the random walk is large.

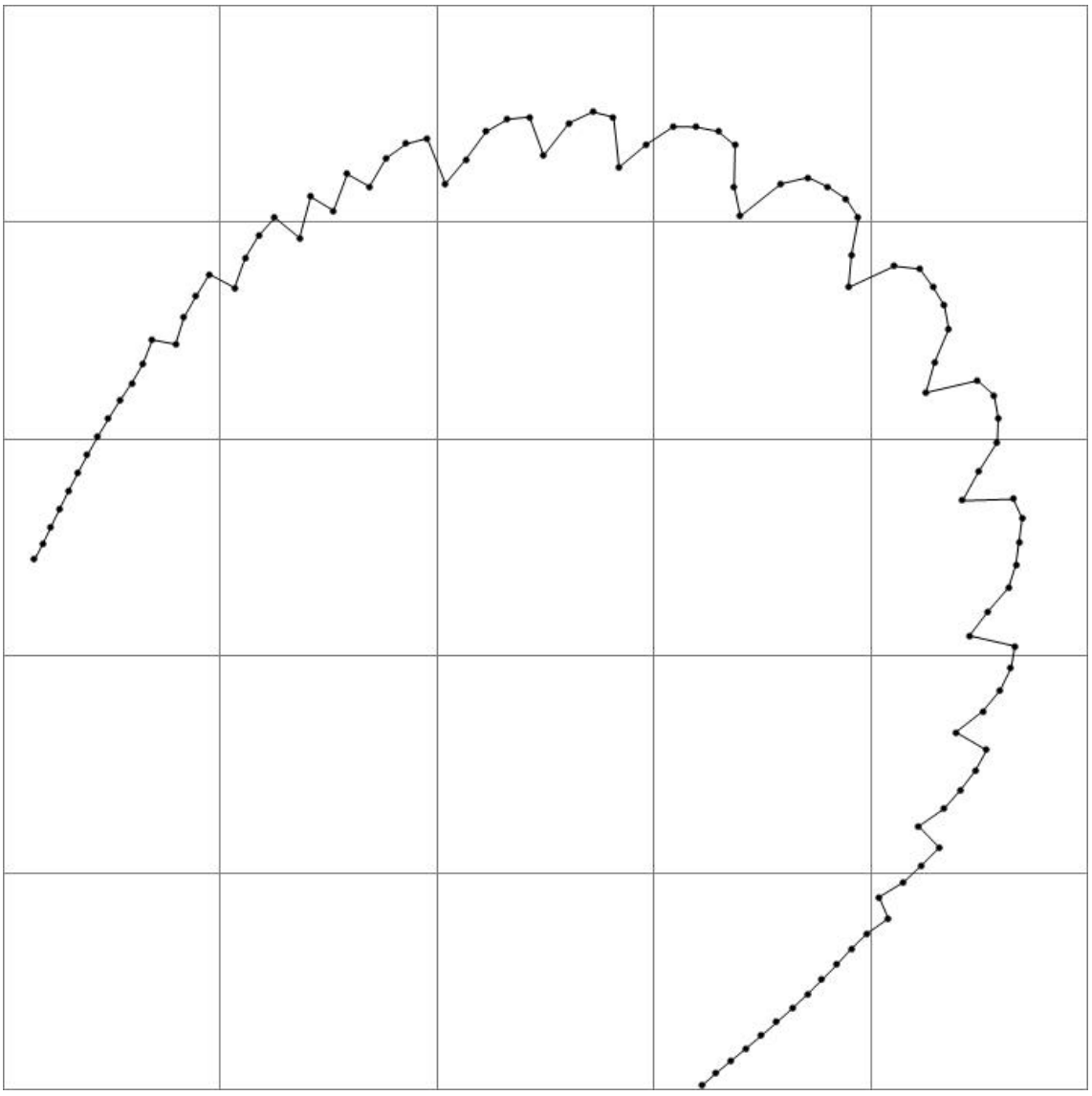

Figure 18: The best fit in 2 dimensions for 100 points generated using a jump of length 1 in a new direction orthogonal to all the previous ones.